MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY OF ALGIERS 2 ABOUELKACEM SAADALLAH

FACULTY OF FOREIGN LANGUAGES

DEPARTMENT OF ENGLISH

An Investigation of the Reliability and Validity of EFL Students' Vocabulary Performance Assessment Using Generalizability Theory

Thesis Submitted to the Department of English in Partial Fulfilment of the Requirements for the Doctorate Degree in English Linguistics and Didactics

Submitted by:
Mrs. Wassila Tebaa

Under the Supervision of:
Dr. Yasmine Boukhedimi

Board of Examiners

Chair: Prof. Samira Arar, University of Algiers 2

Supervisor: Dr. Yasmine Boukhedimi, University of Algiers 2

External examiner: Dr. Amina Hamdoud, ENS Bouzareah, Alger

External examiner: Dr. Nora Achili, University of Boumerdes

Internal examiner: Dr. Fizia Sari Ahmed Bouchama, University of Algiers 2

Internal examiner: Prof. Sihem Bouzar, University of Algiers 2

2025

This page is intentionally left blank

DECLARATION

I hereby declare that the substance of this PhD thesis, **entitled 'An Investigation of the Reliability and Validity of EFL Students' Vocabulary Performance Assessment Using Generalizability Theory'**, is entirely the result of my investigation and has been produced solely by myself under the supervision of Dr. Yasmine Boukhedimi that due reference or acknowledgement is made, whenever necessary, to the work of other researchers. It has not been submitted, in whole or in part, in any previous application for a degree.

Place: Algiers

Date: July 2024

Candidate: Wassila Tebaa

DEDICATION

I dedicate this modest work to my loving parents, Salah and Nouara

To my husband Fateh; to my children, Abderraouf, Anfel, Sarah, and Hadyl; brothers and sisters, to extended family;

Their unending

encouragement has sustained me through the

many years of my studies and research endeavour.

To friends and colleagues who have inspired my

Work,

who remind me every day of what is important.

ACKNOWLEDGMENTS

I praise and thank Allah, The Almighty, The Most Merciful, and The Most Gracious, for His Greatness and Generosity for giving me wisdom, courage and strength to explore the mysteries of Generalizability Theory and thus accomplish this modest piece of research that contributes to a better understanding of the theory in play.

First and foremost, I express my deepest gratitude to Dr. Yasmine Boukhedimi, to whom whatever words I write can never express my acknowledgments for her valuable insights, patience, and fruitful mentorship. Without her expertise and role modeling during my thesis journey, from the initial refinement of my research proposal to the final submission of my thesis, it would never have been completed without her constant support and guidance.

My special gratitude go to the PhD thesis committee members: Dr. Amina Hamdoud, Dr. Fizia Sari Ahmed Bouchama, Dr. Nora Achili, Prof. Samira Arar, and Prof. Sihem Bouzar. Thank you very much for your time and efforts, reading, evaluating and providing valuable feedback on the many pages I wrote.

I am extremely indebted to Farouq Tebaa, a professor at Farhat Abbas University of Setif for his inspiring thoughts and following his guidance for the study statistics.

I would like to thank the many people without whose help the research findings presented in this thesis would never exist. Special thanks go to the inspector Babaissa and Asma Abdelaziz, the secondary school teacher who rated the students' performances, to all the teachers who participated in the process of reviewing the checklist item writing, to Samah Benzarouq, the head of department, and to all those teachers, who facilitated the test administering procedures. To all the students who collaborated either responding to the survey questionnaire or taking the test.

Lastly, and most importantly, thank you to my supportive network of friends, colleagues and family for their moral support.

Abstract

"Assessing the assessment" involves determining the factors affecting its quality, a concern that seems to have been overlooked or insufficiently addressed in the field of language testing. The development of assessment procedures that target vocabulary knowledge and skills of use by first year degree students of English needs more refinement due to sources of measurement error, which in general threaten consistency and accuracy of assessment. This emphasizes the necessity to implement the principles of generalizability (G) theory to gather evidences on the reliability and convergent validity of observed test scores. This study, therefore, was conducted to estimate the consistencies and inconsistencies of students' obtained scores. A descriptive method was used to collect quantitative data in order to identify sources of error in a test taken by 113 students who were newly enrolled in the Department of English at ENSB (Ecole Normale Supérieure de Bouzareah) who sat for a written in-depth productive vocabulary knowledge test. Eight communicative tasks prompting complex situations were constructed to elicit students' competency in applying previously acquired vocabulary knowledge to solve new problems. Two raters scored students' products and the data thus obtained were computed via EduG software package and analyzed through three generalizability (G) and four decision (D) studies. The objective of these D studies was to use the sources of variability determined in the G studies in order to design a measurement procedure that can minimize the magnitude of error variance. The G studies revealed that sources of variability that largely affected reliability of scores were attributed to student-task interaction, student-rater interaction, and the residual component (unmeasured components). Besides, validity was affected by student-task and student-theme interaction. The D studies indicated that the generalizability coefficients were different across study designs and a maximum of five tasks with one rater and four tasks with two raters would yield acceptable levels of generalizability. This study implicates that estimation of measurement precision using G theory is crucial for improvements of assessment methods.

Keywords: performance assessment, generalizability theory, error variance, reliability, convergent validity, vocabulary communicative tasks, EFL students

TABLE OF CONTENTS

CHAPTER TWO: QUALITY CRITERIA FOR PERFORMNCE/COMPETENCE ASSESSMENT

CHAPTER THREE: GENERALIZABILITY THEORY

PART TWO: PRACTICAL CONSIDERATIONS

CHAPTER FOUR: RESEARCH DESIGN AND PROCEDURES OF DATA COLLECTION AND ANALYSIS

CHAPTER SIX: DISCUSSION, INTERPRETATION OF THE FINDINGS AND IMPLICATIONS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

Abbreviation	Definition
ANOVA	Analysis of Variance
AWL	Academic Word List
BAC	Baccalaureate
CEFR	Common European Framework of Reference
CET	College English Test
CTT	Classical Test Theory
DPVKT	Depth of Productive Vocabulary Knowledge Test
D Studies	Decision Studies
DTs	Depth Tests
DVK	Depth of Vocabulary Knowledge
EduG	Software Package
EFL	English as a Foreign Language
ENSB	Ecole Normale Superiure-Bouzareah
ESL	English as a Second Language
Etc.	extera
FL	Foreign Language
FLLs	Foreign Language Learners
G Studies	Generalizability Studies
G Theory	Generalizability Theory
IRT	Item Response Theory
L1	First Language
L2	Second Language
LCT	Listening Comprehension Test
LFP	Lexical Frequency Profile
MCQs	Multiple-choice questions
PVLTs	Productive Vocabulary Levels Tests
SL	Second Language
STs	Size Tests
TL	Target Language
TOEFL	Test of English as a Foreign Language
VKS	Vocabulary knowledge scale
VLTs	Vocabulary Levels Tests
VSTs	Vocabulary Size Tests
RQ	Research Question

INTRODUCTION

Decades ago, the educational literature hinted at a dissatisfaction with traditional assessment. The latter has long been purely based on objective standardized multiple-choice and true-false formats. One could argue that classical assessment is conventionally uni-dimensional, as it was entirely knowledge-based (Linn et al., 1991). It covers decontextualized knowledge and skills and puts more emphasis on randomization of answers (Allem, 2004). Test takers have simply to select from predetermined list of responses, answer rapidly and hazardously, thus eliminating conscious decision and creativity. Along with this, it has been revealed that true-false response formats and short answers applied in many educational institutions do not measure higher-order thinking skills. These test formats, even arguably proved to be valid and reliable, have been questioned for their utility in assessing or targeting performance (Allem, 2000, 2004; Linn et al., 1991).

Classical assessment received severe criticism for its effects on teachers and learning. It measures factual knowledge and stresses discrete skills, promotes artificial short answers, forces students to work individually, teaches to the test, and deprives teachers of opportunities to cover or test various content areas or domains of knowledge (Linn et al., 1991). Such practices emphasize the measurement of lower-order cognitive processes. They favor knowledge which does not reflect teachers' expectations. That is, there is always a mismatch between what is being tested and what is derived from a test. They do not, for example, enhance productive knowledge such as ability of expression or some other kinds of behavior, be it proficiency. In this sense, conventional assessment does not derive accurate decisions about an examinees' success or failure, as test results do not reflect students' true ability. Simply put, it highlights direct assessment that stresses knowledge rather than competence (Allem, 2000, 2004).

Almost without exception, those traditional assessment features characterized vocabulary assessment in the twentieth century. Test takers were required to provide a short dictionary word definition, select the correct word meanings solely from a list of readymade responses, as illustrated in multiple choice formats incorporating odd words, fill in the blank, identify the opposite or synonym when testing knowledge of sense

relations, matching words, …etc. These assessment practices, as it seems, put more focus on either word recognition or meaning retrieval but, absolutely, fail to enhance learners' ability to construct or conceptualize their own meanings (e.g.,. find out a synonym and use it to write correct meaningful sentences or paragraphs). Despite their effectiveness in assessing students' vocabulary size and development, these assessments decontextualize vocabulary knowledge and neglect its deep knowledge aspects such as word use in context. Hence, it is highly important to stress that vocabulary assessment should be more comprehensive or contextualized, as every word has its situational use and communicative stress. It should promote learners' ability to use words in context in order to facilitate communication in every day writing and conversations.

As a reaction to these limitations, an increasing interest in alternative assessment methods based on performance assessment flourished in the educational scene. Educational stakeholders and assessment specialists shifted interest from merely incorporating Classical Test Theory (CTT) to the application of more recent theories of assessment, namely Item Response Theory (IRT) and Generalizability (G) theory. They shifted away, henceforth, from traditional assessment being absolutely based on multiple-choice and short answer tests toward more alternative assessments based on performance (Linn et al., 1991), emphasizing what learners can do with the language knowledge. CTT goes far hand in hand with traditional assessments in estimating the quality of assessment, including validity and reliability. It is typically described by its uni-facetedness and utility of different methods, such as test-retest reliability and inter-rater reliability (see chapter three for a full description) to estimate consistency of test scores. It also depends on many assumptions, among which students' observed score is partitioned into true score and a single source of measurement error. However, it was disapproved for its deficiency to distinguish between several measurement errors, and its defect in the treatment of complex measurement situations. Since, afterwards, performance situations have become complicated and multi-faceted and its facets interact while estimating its assessment quality. This shift of focus urged researchers to look for more measurement methods fitting the requirements of the new tendency in assessment to obtain accurate, valid and reliable scores.

This new insight has become more apparent in many educational systems all over the world ever since. The impedance of developing measurement theories and more recent prevailing assessment methods paved the way to the concept of alternative assessment to emerge in the educational realm. In effect, teaching and learning practices started to adopt performance assessment methods as an alternate solution to traditional assessment. Alternative assessment, as opposed to traditional assessment, enhances students' competences to apply their previously acquired knowledge to solve novel, complex, and realistic problems derived from real, social and professional everyday life needs of students. Performance-based assessment promotes the mastery of complex competences and skills. It seeks to achieve valuable results and positive values in education, and bring about citizens with high self-efficacy, autonomy and motivation to cater for the requirements of the recent complex life situations and recent social changing needs. It is also characterized by its multi-faceted constructs (Norris, 2001), open-mindedness (Chalhoub-Deville, 2001; Herman et al., 1992), realistic situations and open-endedness of questions (Skehan, 1998b), and scoring difficulties (Norris et al., 1998).

Consequently, specialists, researchers, educationists and measurement stakeholders (Baxter et al., 1992; Shavelson et al., 1993; Jonsson & Svingby, 2007; Brennan, 2000; Parkes, 2001) encountered many challenges imposed by alternative assessment requirements. The main concern is to look for methods appropriate for estimating students' performance in an objective, consistent and accurate manner. Performance assessments address many complex issues, such as measuring students' performances in a given domain via engaging examinees in complex situations. Students' performance needs to be rated by means of subjective judgments set up by different raters, as questions are often open-ended and require learners to structure their answers themselves. Students might even perform complex situations and various tasks in distinctive occasions. In essence, performance-based assessment, as it seems, is multi-dimensional. It depends on many facets of measurement or conditions to treat students' performance namely tasks, occasions, raters, contexts …etc. which, when treated, either individually or jointly, should yield consistent ratings.

This impetus urged psychometricians to make use of Generalizability theory to assess complex measurement situations, those that CTT is no longer able to treat and control at a time. With an increasing emphasis on improving performance assessments, successively, the demanding need to adopt Generalizability theory as an appropriate statistical method increased for estimating dependability and validity of behavioral measurements. This theory is used to investigate, treat and control the sources of error variance, affecting score consistency and generalizability. These facets refer mainly to tasks, raters, assessment methods, occasions, teaching methods, scoring methods…etc. and interaction-effects between these facets. Furthermore, in a decision study, generalizability coefficients can be employed to make educational decisions in relation to students' performance scores (relative generalizability coefficients, absolute generalizability coefficients, and criterion-referenced generalizability coefficients). Generalizability theory further displays effective methods that can be deployed to improve reliability of performance assessment scores, a feature that never existed in the CTT (Brennan, 2010).

It follows logically that, advantages of Generalizability theory pushed researchers, such as Brennan (1992a, 1992b, 1992c, 1997), Shavelson, Baxter and Gao (1993), Ruiz-Primo, Baxter and Gao (1993), and Baxter, Shavelson, Herman, Brown and Valadez (1993) in the late 1980's and the 1990's to direct their attention towards Generalizability theory as a useful method to estimating reliability and validity of behavioral measurements. From that time on, many studies were conducted in an attempt to apply Generalizability theory in educational assessment and, even, in other domains of interest like, psychology, biology, economics, technical domains, foreign language (FL) teaching (e.g., Huang, 2009). Generalizability theory increasingly showed prominent results which contributed to the development of measurement methods that sound more accurate and congruent in investigating the dependability of performance assessment scores over several domains, such as language, science, mathematics, art assessments, etc.

Nevertheless, if we consider educational reform in competency assessment, it is highly important to stress that little attention has been paid to investigate the

psychometric properties of students' competency assessment scores, using Generalizability theory despite its application in performance assessment that witnessed great interest, especially after educational reforms occurred in the United States. Education that once emphasized inputs or knowledge, is now, after the reforms has shifted focus towards outputs and processes, moving from the assessment of knowledge to the assessment of learners' ability to apply prior knowledge to solve new problems.

I. Rationale for the Study and Statement of the Problem

Achievement of accountability in teaching entails linking instruction to performance standards and measurable outcomes. This recent trend has led to the need to promote purposeful tasks and performances within educational curricula and research agendas. In higher education, EFL (English as a Foreign Language) students are often expected, to a large extent, to be able to apply their previously acquired knowledge to solve novel real life problems, especially associated with performing academic tasks. Students are required to go far beyond superficial knowledge towards constructing deep meanings (purposeful learning), using and integrating their knowledge to cater for the social complex situations and academic language changing needs and demands.

In the department of English at ENSB (teacher training school known as Ecole Normale Supérieur de Bouzareah), where the current research has been conducted, first year students enter the school with an exist profile composed of a bunch of vocabulary items particularly taught at terminal classes level during secondary education. When referring to the teacher's guide *"New Prospects"* and the ministerial documents accompanying it, one notices that these students are expected to have developed three major competencies throughout the whole course of instruction (Ministry of National Education, 2017). Thus, they are able to interact, interpret and produce in the target language. They are fully capable of interacting with the task situation and instruction, interpret and understand the task content and context upon which they can write at least short constructed responses. Based on the ministry of national education expectations, these students must have developed rich vocabulary repertoire and effective study habits, mainly related to using familiar vocabulary in new contexts. That is, they can use their vocabulary range or store to interact with tasks, interpret tasks content and

produce certain written messages related to different topics. This implies new Baccalaureate (BAC) holders' ability to be engaged in an attainment test where they are required to use a number of target words taught during their third/last year high school course to write short constructed responses or short paragraphs related to a given context. These students are also expected to demonstrate their ability to use vocabulary acquired to solve new problems set up in different authentic situations designed in accordance with the textbook themes.

As a teacher researcher, I have developed an interest in researching vocabulary because I found lexical knowledge[1] to be significantly determinant of learners' language proficiency and academic success in reading comprehension, writing production, listening comprehension and speaking production in the FL and L2 (Second Language). Although many empirical studies have established predetermined vocabulary size and lexical coverage targets for specific language ability (e.g., Hazenberg & Hulstijn, 1996; Hirsh & Nation, 1992; Hu & Nation, 2000; Laufer, 1989, 1992, Thékes, 2018), little research has yet addressed such issues as word knowledge and use in written discourse contexts or, more precisely, within task-based paradigms. The current research is, thus, motivated by lack of empirical evidence on the assessment of vocabulary knowledge in context rather than in isolation as done mostly by vocabulary size and depth tests, a theme treated in Chapter 1.

My interest does not cover testing students' vocabulary knowledge only, but extends to involve an investigation of sources of error that might be affecting measurement precision and consistency of scores obtained from an assessment of vocabulary knowledge. Again, as a teacher researcher, I have always been concerned with the issue of "washback" effect. To FL students, scores obtained from different modules are often questionable. Students' peers used to compare their performance scores with others, particularly in tests with open-ended formats, such as essays, where students respond differently. Do inconsistencies of test scores result from teachers' bias or subjectivity? Are there any other features intruding within the process of measurement and scoring? Are there any interfering issues associated with the test format, length and situations or with test items and instructions? Is this doubt typically

related to the students' performance and conditions of passing the test? Are there any other features affecting measurement situations that I do neither notice nor expect? However, in some tests I took years ago, like MCQs (multiple-choice questions), I never raised doubt about score consistency. I finally could argue that it is salient that most of measurement procedures are not perfect and are, therefore, often subject to measurement error (Yashim et al., 2021). Sources of measurement error must be the direct cause of doubt and untrustworthiness in the validity and reliability of tests and of observed scores henceforth. My concern, then, is to estimate the sources of error that might affect the validity (accuracy) and reliability (consistency) of vocabulary observed scores gained from students' vocabulary performance assessment.

The current research is based on the assumption that vocabulary performance assessment is affected by various sources of variability such as tasks, raters and themes (word choice and task thematic orientation) as systematic errors, and other indefinable unsystematic errors of measurement that might speculatively emerge within the measurement situation. Vocabulary assessment is mostly an incremental and challenging obstacle EFL teachers must overcome. Essentially, uncertainty is always attached to school and students' scores as a result of inconsistency of achievement that seems to be a prominent threat to reliability and validity (Yashim et al., 2021), especially within performance-based assessment paradigm where open-ended questions are appealing and, thus, for structured response types. Hence, the need to examine sources of variation and errors, especially those stemming from inconsistency across raters, modes or tasks and themes become apparent. In this context, Elliot and Roach (2007) state that "With the popularity of alternative assessments (e.g., portfolios, performance tasks) increasing over the past 10 years, (…), researchers must determine the precision of the scores associated with these assessments" (p.372). These researchers indicate that investigating the dependability of scores obtained from performance assessments is of crucial importance so that researchers can establish more reliable and valid assessments upon which accurate interpretations of test results would be made to further make decisions on students' academic success or failure.

Influences on vocabulary tests and inconsistencies of vocabulary test scores, in particular, can be explained by the difficulty to model and assess vocabulary knowledge due to the incremental nature of vocabulary knowledge itself. Teachers consider vocabulary as one of the significant subjects taught in EFL schools, eventhough this significance has not yet been supplemented by effective assessment. The difficulty to model and measure vocabulary knowledge is an issue related totally to the interference of factors like, determining pre-existing vocabulary knowledge, the selection of vocabulary items, and what should be involved in testing among the various components of knowing a linguistic item. What words to test, including unknown words in measurement instruments, are other critical issues when modelling and assessing word knowledge.

Modeling and assessing EFL students' vocabulary knowledge may influence students' observed test scores. As such vocabulary performance, like any other performance assessments, is affected either by systematic deviations that may refer to rater bias or rater drift, tasks or test items, test format, and thematic orientation (word selection, content coverage and context), or unsystematic deviations associated, for instance, with students' background knowledge. Besides its assessment, vocabulary acquisition and instruction are more complex and challenging for FL learners/teachers. Azfal (2019) confirmed that EFL learners encounter many vocabulary learning difficulties even at the university level, and identified these as: meanings of new words, spelling, pronunciation, correct use of words, and guessing meaning from context. New word learning deficiencies lead to limit students' vocabulary knowledge that negatively affect students' reading comprehension, writing production, listening skills and understanding and, above all, overall communicative competence. As a result, it seems that the major sources of error affecting vocabulary measurement are related to tasks (difficulty to develop a model test), themes (word selection mysteries put within an organized set of ideas relevant to a given theme), and raters being a source of variability in performance assessment providing subjective judgements.

Accordingly, the core issue of this research revolves around the sources of error that might affect the vocabulary measurement precision within the context of

performance-based assessment, especially that Hathcoat and Penn (2012) have suggested that one more problem associated with the reoccurring theme is that "Little research has examined the consistency of scores across multiple authentic assignments or the implications of this source of error on the generalizability of assessment results" (p.16). Unlike performance assessment, objective standardized multiple-choice vocabulary tests are not subject to various variance components as questions used are precise and close-ended, and thus are not opened to different interpretations. They are based on selection of answers from convenient lists. They are not influenced by rating deviations as scoring is easy and consistent.

In line with the above discussions, it seems that exploring vocabulary knowledge assessment might be challenged by two major obstacles: one issue concerns its modeling and measurement caused by the multidimensionality of vocabulary that might affect students' performance, the second is related to sources of variation that may negatively affect students' observed scores. Consequently, the present study focuses on investigating the relative effects of tasks, raters and themes on the dependability and generalizability of students' scores obtained from an assessment of vocabulary knowledge and use. It is an attempt to estimate the quality of the present measurement instrument, namely its psychometric conditions of validity and reliability.

II. Objectives of the Study

The overarching aim of the current study, therefore, is to estimate the reliability and validity of scores obtained from a vocabulary performance assessment administered to newly enrolled students at ENSB through the application of Generalizability Theory, which will be used as a methodological tool to guide the estimation of score reliability and inform assessment design.

More specifically, the study seeks to investigate the extent to which various sources of variance namely tasks, raters, and themes, impact the reliability and generalizability of observed scores. This includes examining how modifying the number of tasks, raters, or themes affects the precision and consistency of test results. By identifying and analyzing these variance components, the study aims to determine

optimal assessment conditions that can enhance score reliability while maintaining test efficiency.

In this context, understanding the impact of different facets on measurement error allows for informed decisions regarding test design. For example, psychometricians may reduce the number of raters from five to three or the number of tasks from ten to six to improve efficiency without compromising reliability. Such decisions are grounded in pre-estimated variance components that help to maximize the generalizability coefficient.

Furthermore, this study aims to evaluate the validity of the inferences drawn from the assessment, particularly the extent to which these inferences are generalizable to broader contexts. By doing so, the research contributes to a more comprehensive understanding of the construct validity of vocabulary performance assessments.

An additional goal of this study is to derive pedagogical implications based on the application of Generalizability Theory. These implications are expected to inform researchers, EFL educators, and assessment specialists in refining practices for designing and scoring vocabulary assessments, especially those that evaluate learners' ability to complete complex, communicative tasks involving productive vocabulary use.

In order to address these aims, the study is guided by a number of research questions that are set forth in the upcoming section.

III. Research Questions

Vocabulary assessment, particularly in productive contexts, is inherently multifaceted. Variability may arise from several sources, including the types of tasks used to elicit vocabulary knowledge, the raters who evaluate performance, and the thematic content of the assessment prompts. To ensure that test scores are both reliable and valid indicators of learners' vocabulary proficiency, it is essential to identify and quantify the impact of these factors.

Generalizability theory provides a powerful analytical framework for this purpose, as it enables the estimation of multiple sources of measurement error simultaneously. This study applies G Theory to explore the extent to which tasks, raters, and themes,

key test design facets, contribute to score variability and to determine optimal conditions for achieving reliable and valid vocabulary assessments.

Accordingly, the study is guided by the following research questions that are designed to examine their influence on test score reliability and validity through the lens of Generalizability Theory:

1- What is the relative effect of tasks and raters on the generalizability (for relative decisions) and dependability (for absolute decisions) of scores obtained from a vocabulary performance test?

2- What is the relative effect of tasks and themes on the generalizability and dependability of scores obtained from a vocabulary performance test?

3- What is the relative effect of tasks, raters and themes on the generalizability and dependability of scores obtained from a vocabulary performance test?

4- What is the effect of decreasing the number of tasks designed to assess vocabulary performance on the generalizability and dependability of test scores?

5- What is the effect of decreasing the number of raters on the generalizability and dependability of test scores?

6- What is the relative effect of tasks, raters, and themes on the construct validity of a vocabulary performance assessment?

By addressing these questions, the study aims to provide practical, evidence-based guidance for the development of vocabulary assessments that are both psychometrically robust and pedagogically meaningful.

Since the purpose of any test, amongst language tests, is to collect data on students' performance in a given domain, here word repository and linguistic repertoire, we aim to investigate the quality of performance assessment systems. The qualities underlying the type of assessment encompass the criteria of validity and reliability of students' performance scores. The main aim, then, is to examine the reliability of assessment procedures. That is, the effects that the different tasks, scorers and themes have on the score dependability focusing on sources of measurement error and score reliability.

IV. Significance of the Study

No one can deny the importance of word knowledge in learning a foreign or a second language. Grammar and pronunciation are two other important language components too, but they are considered in a second order of importance, because individuals can narrowly communicate with incorrect grammar and mispronounced words, but can never communicate with no words. In this context, Wilkins (1972) states that "without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (p. 111). This quote undeniably highlights the importance and place of lexical knowledge in FL components. In fact, it is far more impossible to communicate without vocabulary, since it is determinant of learners' academic success and achievement in listening, reading, writing, speaking and overall language proficiency (Qian, 1999; 2002; 2004).

Since evaluation is considered a vital element in the teaching learning processes and is looked upon as an inevitable tool, evaluating its data is of great value. That is, one way to amend and improve the curriculum is to improve assessment and other materials employed in the program. Evaluation seeks to identify strengths and weaknesses, so that optimum use can be made of strong points and weaker points can be adapted or substituted from other sources. Evaluation of student learning is, subsequently, an essential feature of the teaching learning processes and that of pedagogical activity.

Elaborating assessment procedures and properties is, by no means, achieved by assessing the assessment, a significant feature to the present study. Generalizability theory is a key foundation stone towards the assessment of psychometric properties of test scores obtained from the current productive vocabulary knowledge test.

V. Operational Definitions of Research Key Concepts

The following key terms are typical and will be recurrent throughout our research. They are operationalized within the context of the present study. This research work focuses on the application of Generalizability theory to estimate the psychometric properties, namely validity and reliability of a productive vocabulary knowledge performance assessment. Test takers' scores obtained from an attainment test are to be

analyzed using ANOVA and EduG; the two procedures will be used in the analysis of variance to estimate the sources of error variance (facets of measurements) that might be affecting the present measurement precision and score generalizability.

Generalizability theory: A conceptual framework and statistical multifaceted method used for assessing the consistency of test scores across various facets of measurement, such as items, tasks, raters, occasions, persons, modes, scoring methods, teaching methods, … etc.

Generalizability: The extent to which test scores can be reliably generalized across facets of measurement. In other words, how well observed scores reflect the true performance of a students across the universe of possible testing situations.

Validity: A valid test measures what it is supposed to measure. For example, a test designed to measure productive vocabulary knowledge should measure students' ability to use (production) words in context rather than to test their receptive knowledge (meaning recognition). In this particular research, validity entails the appropriateness and trustworthiness of test scores upon which inferences and decisions will be made. Validity, here, is taken to mean a non-test property, but an inference property.

Construct validity: Construct validity in this study refers to the degree to which the vocabulary performance test accurately measures the intended construct of productive vocabulary ability. This includes minimizing construct-irrelevant variance introduced by raters, tasks, and themes, as assessed through Generalizability Theory.

Reliability: A reliable test is consistent and dependable. If you administer the same test to the same student or matched students on two different occasions, the test will yield almost similar results. Reliability, however, might be affected by test unreliability caused by test item length and structure or rater unreliability due to human error and subjectivity. The terms reliability and depeddability wil be used interchangeably throughout this work.

Productive vocabulary knowledge: The term refers to the test measuring students' ability to use target words in appropriate contexts. Productive vocabulary is

used interchangeably with active vocabulary corresponding to the writing skill as vocabulary knowledge is tested via writing.

Performance assessment: Performance assessment refers to any purposeful task performed by students in a specific occasion examined by an expert rater who considers both the process of completing the task and its product. In the current context, it is not based on a paper-and-pencil selective response rather on written production elicited via open-ended questions or performance tasks where students can demonstrate their productive vocabulary knowledge and ability to use target words in relation to specific discursive and thematic contexts.

Attainment test: It is meant to measure how much students have learned in a specific content area after a course of instruction. In this study, it aims to measure students' productive depth of vocabulary knowledge, namely knowledge of word meanings, knowledge of word forms and word use in context.

ANOVA: This acronym stands for Analysis of variance, a procedure in Generalizability studies used for estimating the variance components of measurement in a specific testing situation.

EduG: A computer software program used for data analysis to estimate not only the variance components for the main and interaction effects for examinees, tasks, raters, and themes but also the generalizability coefficients.

This research work is descriptive. It applies Generalizability theory to identify sources of error variance and their relative effects on score reliability and validity.

The last section of this introduction lays out the structure of the research study and provides a brief description of the contents of each chapter.

VI. Structure of the Thesis

The material in this thesis is structured into six different but related chapters. The theoretical part consists of the first three chapters devoted to the review of the related literature, and the practical part, in turn, is composed of the other three remaining chapters about the research fieldwork.

The First Chapter outlines the extensive literature on vocabulary assessment. It focuses on a specification of vocabulary knowledge operational definitions, the attribute under study. It explains the significance of vocabulary knowledge to the four language skills and overall language proficiency. The chapter involves an overview of depth of word knowledge conceptualizations, vocabulary test types, previous research work conducted on the subject, factors affecting validity of vocabulary tests, and procedures of vocabulary test design and development.

The Second Chapter deals with performance-based assessment principles as opposed to traditional assessment. It includes an overview of performance and performance assessment characteristics, major quality criteria used for "assessing the assessment", including validity as a key feature in the study, scoring performance, and presenting sources of error in performance assessments. Adding this chapter to the literature is justified by a desperate need to have an extensive review of literatures on the quality criteria necessary for test validation, development and use. It emphasizes modern properties of test scores, reliability and validity set up in performance assessment standards.

The Third Chapter considers the theoretical underpinnings of Generalizability theory. It conceptualizes and discusses the evolution and advantages of the psychometric Generalizability theory in connection with the limitations of Classical Test Theory. The chapter also explores the major concepts and principles of Generalizability theory and reviews earlier research related to the theme in question, and finally highlights its contribution to the research paradigm.

The Fourth Chapter, which opens up the practical part, tackles issues that often arise in the research design. It describes the research method, data gathering procedures and research instruments, sampling, test design and development (purpose, piloting, and administering), developing scoring guide, and finally applying Generalizability theory and procedures for data analysis.

The Fifth Chapter reports the research findings obtained from the current measurement procedure implemented to quantify observation.

The Sixth Chapter, a follow up, that discusses and interprets the major research findings. This closing chapter also provides assessment specialists with a number of suggestions for implementing Generalizability theory principles to determine the quality of vocabulary performance assessments.

The thesis ends up with a general conclusion and further research potentials and points to some study limitations. For referencing, the thesis includes appendices and a reference list at the end.

CHAPTER ONE

ASSESSING VOCABULARY KNOWLEDGE

Introduction

> *Words, so innocent and powerless as they are, as standing in a dictionary; how potent for good and evil they become in the hands of one who knows how to choose and combine them.*
>
> *—Nathaniel Hawthorne (1804-1864).*
>
> From: *https://www.azquotes.com/quote/127016.*

The aforementioned quote, as it seems, contains the most appropriate words ever to introduce this chapter as it encapsulates the power of words, and the central importance of vocabulary henceforth, that is deeply recognized by fluent speakers, literary writers even mostly readers. In this chapter, we will explore vocabulary knowledge as a multifaceted research construct together with issues related to its assessment and testing. Firstly, and most importantly, this chapter sketches out the significance of lexical knowledge to the four language skills, to success in academic performance/achievement, to enrich total vocabulary repertoire and to contribute to the overall language proficiency/ability. It also explores the complex nature of word knowledge and examines how its intricacies make the processes of defining vocabulary knowledge and designing its assessment difficult and challenging.

This chapter further explores the worldwide different approaches applied to measure vocabulary knowledge, and deep word knowledge in particular. It mainly discusses the incremental nature of depth of lexical knowledge dimension and, gives a detailed account of every depth measure. It attempts to explain the concepts utilized throughout the work and notably, considers an explanation of why depth of vocabulary knowledge should be measured productively via writting. Additionally, the chapter emphasizes different practical working terms used in measuring vocabulary knowledge including contextualization. It further argues for test purpose and what target words to measure, and accounts for what aspects of vocabulary knowledge to measure and how to elicit this type of knowledge. Finally, the chapter explores an FL vocabulary

assessment framework that forms the basis for the current research test design and development and concludes with a summary on multiple issues linked to the complexity of lexical knowledge and its modelling and assessment.

1.1. Contribution of Vocabulary to the Four Language Skills and Overall Language Proficiency

Vocabulary/lexical knowledge is widely recognized as a foundational component of foreign and second language (FL/L2) proficiency, with a significant impact on learners' performance across the four primary language skills, reading, writing, listening, and speaking, as well as overall communicative competence. Wilkins (1972) famously stated, "without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (p. 111), emphasizing the essential role of vocabulary. Research suggests that vocabulary breadth (number of words known) and depth (quality of knowledge) are strong predictors of skill performance, particularly when assessed in combination (Milton, 2013).

1.1.1. Contribution of Vocabulary to Reading

Reading comprehension is closely linked to vocabulary knowledge. Laufer (1989) found that learners must know at least 95% of the words in a text to ensure adequate comprehension, while Nation (2001, 2006) recommended 98% coverage, equating to 8,000–9,000 word families for authentic texts like newspapers and novels. Dale (1965) also noted high correlations between vocabulary and reading ability. Qian (1999) reported significant correlations between vocabulary size ($r = .78$), depth ($r = .82$), and reading comprehension ($r = .64$), with depth adding 11% of predictive power over the 3% explained by size. Qian (2002) confirmed these results, showing correlations of .82 (depth) and .78 (size) with TOEFL reading scores.

Ehsanzadeh (2012) found that vocabulary depth ($r = .65$) was more effective than breadth ($r = .50$) in lexical inference and long-term retention. Similarly, Li and Kirby (2014) found stronger correlations for breadth in reading comprehension, while depth was more relevant for summary writing. Swart et al. (2016) demonstrated that higher scores in written and oral vocabulary breadth and depth were associated with better reading comprehension. Zhang and Yang (2016) reported strong correlations for

vocabulary size ($r = .748$, $p < .001$) and depth ($r = .720$, $p < .001$). Karakoça and Köse (2017) identified a moderate positive correlation between receptive vocabulary and reading performance ($r = .429$).

1.1.2. Contribution of Vocabulary to Writing

Writing, a productive skill, requires active use of passive lexical knowledge. Laufer (2013) emphasized the reciprocal relationship between vocabulary and writing: richer vocabulary enhances writing quality, and writing, in turn, facilitates vocabulary development. Engber (1995) found a significant negative correlation ($r = –.43$, $p < .01$) between lexical error rate and writing quality, suggesting that more diverse and accurate vocabulary usage improves essay scoring.

Webb (2009) compared receptive and productive vocabulary instruction. Participants who learned vocabulary productively used 42% of target words in writing, versus 29% for those with receptive instruction. Nation (2006) observed that the most frequent 1,000 word families account for 81% of written texts. Karakoça and Köse (2017) found a moderate positive correlation between productive vocabulary and writing performance ($r = .378$), indicating that vocabulary richness strongly predicts writing fluency and coherence.

1.1.3. Contribution of Vocabulary to Listening

Vocabulary also plays a crucial role in listening comprehension. Stæhr (2009) reported significant correlations between vocabulary breadth ($r = .70$), depth ($r = .65$), and listening scores, with both dimensions explaining 51% of variance. Additionally, Teng (2014) found a very high correlation ($r = .91$) between depth of vocabulary and listening comprehension, with depth having greater predictive power than breadth. Feng (2016) similarly found strong correlations between listening and vocabulary depth ($r = .75$) and breadth ($r = .70$), supporting the necessity of lexical coverage for spoken input.

1.1.4. Contribution of Vocabulary to Speaking

Vocabulary size and depth are integral to spoken fluency and lexical accuracy. Nation (2006) stated that around 2,000 word families are needed for basic conversation, while 6,000–7,000 are required for 98% coverage of spoken texts. Koizumi and In’nami

(2013) concluded that vocabulary knowledge predicts 84% of L2 speaking proficiency. Vocabulary size and depth contributed 63% and 60%, respectively, to speaking fluency, while access speed accounted for only 28%, highlighting vocabulary's primary role in speech production.

1.1.5. Contribution of Vocabulary to Overall Language Proficiency

Vocabulary is also a significant indicator of general language proficiency. Read (2000) argued that proficient learners have larger semantic networks. Schmitt and Meara (1997) observed that suffix knowledge facilitates vocabulary expansion by improving access to word families. Karakoça and Köse (2017) found a moderate correlation between passive vocabulary and general language ability (r = .650), and a strong correlation for active vocabulary (r = .826), confirming that productive vocabulary knowledge is more predictive of language proficiency than receptive knowledge.

In summary, vocabulary knowledge, both breadth and depth, profoundly influences learners' reading, writing, listening, and speaking abilities and is a decisive factor in overall language proficiency. Its role in academic performance and communicative competence underscores its centrality in FL and L2 instruction.

1.2. Vocabulary Knowledge Conceptualization: A Historical Account

The study of vocabulary knowledge has been a focal point in L2 acquisition for several decades, aiming to answer the fundamental question, "What does it mean to know a word?" Various researchers have proposed multiple frameworks over the years, each contributing to our understanding of vocabulary knowledge as a multifaceted and incremental construct. Early works, such as those by Cronbach (1942), Dale (1965), and Richards (1976), laid the foundation for this exploration, presenting overlapping models that emphasized different dimensions of word knowledge.

In Cronbach's (1942) pioneering framework, five types of vocabulary knowledge were outlined. These included: (1) *generalization*, which refers to the learner's ability to define a word; (2) *application*, which refers to how well learners use words in context; (3) *breadth of meaning*, which pertains to a word's multiple meanings across

different contexts; (4) *precision of meaning*, which describes the ability to apply word meanings accurately in various situations; and (5) *availability*, which refers to how easily learners can access and use words. However, this model was critiqued by researchers like Qian (2002), who argued that it underemphasized the importance of morphological and collocational knowledge.

Dale (1965) proposed a four-stage continuum of vocabulary growth, emphasizing the incremental nature of vocabulary acquisition. The stages included: (1) no prior exposure to the word, (2) awareness of the word's existence without understanding its meaning, (3) vague understanding of the word's meaning in context, and (4) full recognition and ability to recall the word in any context. His model highlighted the idea that vocabulary knowledge increases progressively over time.

Richards (1976) further developed vocabulary knowledge conceptualizations by incorporating features like *register*, *word frequency*, and *morpho-syntactic properties*. His framework emphasized that word learning is a complex, ongoing process, involving not only semantic aspects but also recognition of word forms and their various meanings. He identified eight assumptions that outlined the different kinds of knowledge a learner needs to fully master a word, which are further explored in section 1.3.

In the 1980s, Anderson and Freebody (1981) distinguished between *breadth* and *depth*[2] of vocabulary knowledge, offering a dual framework that became widely accepted in later studies (Qian, 1999; Read, 2000; Schmitt, 2014). Breadth refers to the number of words a learner knows to some extent, while depth refers to the quality of understanding of those words. As Anderson and Freebody (1981) stated, depth of understanding implies knowing a word well enough to understand all the distinctions that would be clear to an ordinary adult speaker.

Similar to Dale (1965), Beck et al. (1987) proposed a five-stage continuum of word knowledge, from zero knowledge to full mastery, with the final stage involving not only full recognition but also an ability to use words in complex contexts, including figurative language. Nation (1990) and later (2001, 2013) refined this by developing a comprehensive framework that highlighted form, meaning, and use as essential aspects

of vocabulary knowledge. This framework distinguished between receptive vocabulary (understanding words when heard or read), and productive vocabulary (using words actively in speech or writing).

Henriksen (1999) further expanded on these ideas by proposing a three-dimensional model of lexical competence, which included: (1) *partial-to-precise knowledge* (knowledge of word form-meaning connections), which aligns with vocabulary size; (2) *depth* (a rich representation of word meaning, including sense relations, collocations, and syntactic properties); and (3) the distinction between *receptive and productive knowledge*, where receptive knowledge refers to passive understanding, and productive knowledge refers to active use of words.

Qian (2002) developed a more integrated framework, incorporating aspects of prior models to propose four interrelated components of vocabulary knowledge: (1) *breadth*, or the number of words known; (2) *depth*, encompassing phonemic, graphemic, morphological, syntactic, and semantic knowledge; (3) *lexical organization*, which refers to how words are mentally stored and organized; and (4) *automaticity*, which refers to the ease with which words can be accessed and used in both receptive and productive contexts. This last aspect, automaticity, emphasizes how quickly learners can retrieve and apply words, a critical factor in real-time language use.

Meara (2005) further refined the conceptualization of lexical competence by introducing *vocabulary size*, *depth of vocabulary knowledge*, and *accessibility*. Accessibility refers to how easily learners can manipulate the words they know, a concept similar to Qian's (2002) automaticity and Henriksen's (1999) focus on the transition from receptive to productive knowledge. These models stress that vocabulary knowledge is not just about knowing a word's meaning, but also how well it is structured in the mental lexicon and how easily it can be retrieved for productive use.

In their study, Daller et al. (2007) expanded the concept by introducing a new dimension of *fluency*, which describes how quickly and efficiently learners can use their vocabulary in communicative contexts. This model represents vocabulary knowledge in a *lexical space* with three axes: *breadth*, *depth*, and *fluency*. Breadth is positioned along the horizontal axis, indicating the number of words known; depth is along the

vertical axis, reflecting the richness of understanding; and fluency is represented as the third dimension, showing how readily and automatically learners can access and use the words they know.

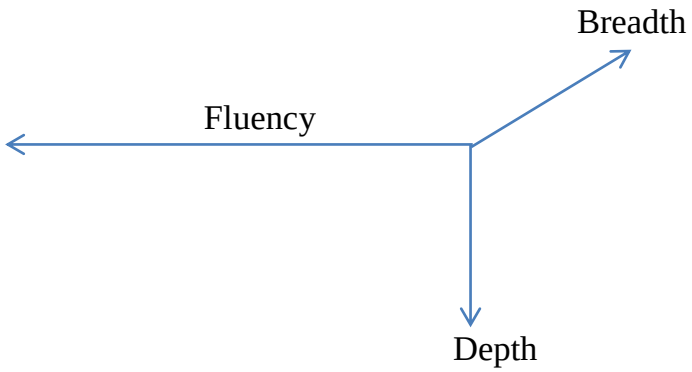


Figure 1.1: The Lexical Space of Word Knowledge and Ability (Daller, et al., 2007, p .8)

The concept of fluency relates closely to Meara's (1996, 2005) work on vocabulary accessibility, and it has become an important addition to the understanding of vocabulary knowledge in communicative situations. Daller et al. (2007) argue that fluency is essential for both receptive and productive vocabulary knowledge, as it involves the ability to recognize and process words quickly, as well as to recall and use them productively in speaking and writing.

Bravo and Cervetti (2008) introduced a continuum of word knowledge specific to content area vocabulary, proposing three stages: unknown, acquainted, and established. In the first level, a learner has never encountered the word, in the second level, a learner recognizes the word and can provide a basic definition, and in the final level, a student can use the word accurately in context. This continuum aligns with the incremental nature of vocabulary knowledge development.

While many frameworks agree on the key dimensions of breadth and depth, they differ in how these dimensions are conceptualized and assessed. For instance, Nation’s (2001) framework includes form, meaning, and use as central components of depth, whereas Henriksen (1999) includes sense relations (e.g., synonyms, antonyms) and collocational restrictions. Other frameworks, such as those proposed by Qian (2002) and Ebrahimi (2017), emphasize the productive use of vocabulary and how depth is

intertwined with automaticity and lexical fluency.

In summary, the development of vocabulary knowledge frameworks highlights that lexical competence is a multi-dimensional, incremental construct. The dimensions of breadth (word quantity) and depth (word quality) are commonly recognized as fundamental, but fluency, lexical access, and automaticity are also crucial for effective word use in communicative contexts. Vocabulary knowledge involves not only knowing a word's meaning but also its form, syntactic properties, usage, and the ability to retrieve and use it efficiently across different contexts. This conceptualization of lexical knowledge forms the foundation for the current research, which focuses on the productive depth of vocabulary knowledge, particularly in terms of word meaning, form, and use in communication.

1.3. Operationalization of Deep Word Knowledge

Vocabulary or vocabulary knowledge, as it has emerged in the above section, is a multi-dimensional theoretical construct which involves various aspects in knowing a word (Cronbach, 1942; Dale, 1965; Richards, 1976 Anderson & Freebody, 1981; Nation, 1990, 2001, 2013; Read, 2000; Daller et al., 2007; Webb, 2013). In vocabulary knowledge literature, a number of conceptualizations have been suggested amongst the famous is the distinction made between breadth and depth.

The notion of depth of knowledge was first introduced by Anderson and Freebody (1981) as a distinct dimension of vocabulary knowledge apart from breadth of knowledge. The former type of knowledge alludes to how well (the quality) a learner knows a word, whereas the second type of knowledge refers to how many (the number) words are known (Anderson & Freebody 1981; Read, 1989, 2000; Qian, 1998, 1999; Qian & Schedl, 2004; Webb, 2013; Schmitt, 2014). In this dichotomy, breadth definition sounds quite accurate since it accounts for the relationship between form and meaning in a numeric sense as illustrated in the case of *Vocabulary Levels Tests* (Nation, 1990). Similarly, when considering the previous definition of deep word knowledge, it seems a bit clear, but when its aspects are taken up it seems rather superficial and in this case its definition differs and becomes much fuzzier. Accordingly, "there is no definition of vocabulary depth that is widely agreed upon" (Webb, 2013, p.1) especially

that aspects involved in depth of knowledge and the incremental nature of vocabulary made a convention on depth impossible. Despite this many attempts have been made to identify depth of vocabulary knowledge and its various components.

One way to conceive of depth of vocabulary knowledge in the research literature is knowing multiple word knowledge aspects. A rampant view of depth lies in dividing it into distinctive features (Schmitt, 2014). Knowledge of a word is complex and multi-componential (Nation, 1990, 2001, 2013). These multi-dimensional aspects extend word meanings, a uni-dimensional component of breadth, to involve semantic relationships, collocations and syntactic behaviours of a word (Bardakçı, 2016).

Long ago, Richards (1976) provided an initial framework on depth of vocabulary knowledge, set up in eight assumptions of different kinds of lexical knowledge a native speaker should develop. In his proposition, he stresses that word learning is an ongoing process in which he itemized the various types of knowledge that are necessary to have a complete knowledge of a word. They are introduced in terms of eight assumptions:

1. The native speaker of a language continues to expand his vocabulary in adulthood, whereas there is comparatively little development of syntax in adult life;

2. Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words we also know the sort of words most likely to be found associated with the word;

3. Knowing a word implies knowing the limitations imposed on the use of the word according to variations of function and situation;

4. Knowing a word means knowing the syntactic behavior associated with that word;

5. Knowing a word entails knowledge of the underlying form of a word and the derivations that can be made from it;

6. Knowing a word entails knowledge of the network of associations between that word and other words in language;

7. Knowing a word means knowing the semantic value of a word; and

8. Knowing a word means knowing many of the different meanings associated with the

word. (p. 83)

Richard's framework emphasizes the complex and multi-faceted nature of vocabulary knowledge (Read, 2000) as it involved more word properties like register and frequency level exceeding meaning recognition of word forms.

Henriksen (1999) in her study on depth of vocabulary knowledge considers three dimensions of vocabulary development: *breadth, depth, and receptive and productive knowledge*. Henriksen's definition of vocabulary depth is equated with that of Richards' basic assumptions or aspects of word knowledge and to that of Nation's (1990) components involved in knowing a lexical unit (see Table 4, p.110). She declares that deep word knowledge, or what she terms "rich meaning" (p. 306) calls for meaning recognition in addition to other paradigmatic (synonymy, hyponymy), syntagmatic (collocations) sense relations, syntactic patterning and morphological features of a word.

Following this path, Read (2000) contends that Henriksen's theoretical conception of depth is said to represent the best conceptualization of quality of word knowledge particularly in that it provides the specification of what construct dimensions a researcher on vocabulary should measure in her/his particular studies. Yet, he (2004) suggests a distinct conceptualization of depth in L2 vocabulary acquisition.

The author reviewed research on vocabulary knowledge and assessment and concluded that depth of vocabulary knowledge is conceptualized under three headings (dimensions): *precision of meaning*, *comprehensive word knowledge*, and *network knowledge*. These three major approaches to Read's conception of vocabulary knowledge is now examined.

1. *Precision of meaning:* Read's precision of meaning is equated with Anderson and Freebody's (1981) definition of depth of vocabulary knowledge. In this first conception of depth Read distinguishes between types of knowledge of meaning a learner has; a learners' meaning recognition ranges from having narrow or imprecise thought of a word to a more elaborated and precise understanding of its meaning. To illustrate, Polysemous words are inherently vague in that they carry various meanings to the extent that a learner may have a limited or general idea about especially out of

context (e, g. the term *'book'* is contextually dependent which means a set of printed pages in one situation and making a reservation in hotel in another); and specialized or technical meanings of which a learner has accurate and advanced recognition (e.g., a learner enrolled in specialised linguistics courses precisely recognizes words like morpheme, paradigm, syntagm; and a learner of literature would perfectly know words like prosody, stanza, rhyme).

2. *Comprehensive word knowledge*: Although most of comprehensive word knowledge aspects are listed in Cronbach's (1942) and Richards' (1976) frameworks, Read equates this approach with Nation's (2001) conception of what is involved in knowing a word. In this sense, comprehensive word knowledge requires more than grasping the meaning component to include other components of form (pronunciation, spelling, word parts); meaning (form-meaning relationship, concept and referents, associations); and use (grammatical functions, collocations, constraints on use i.e. register and frequency). From an assessment perspective, Read asserts that researchers' attempt to measure all the comprehensive word knowledge facets yield in completely perplexed test design.

3. *Network knowledge*: It represents a network approach, which is based on the assumption that when a learner's vocabulary size increases, a learner incorporates newly learned words into a connected lexical network settled in the mental lexicon. An L2 learner develops the ability to accumulate the newly acquired words to his/her existing knowledge of words, and distinguishes how words are semantically different or related to other words and how to connect these words in a language. This dimension sounds similar to Meara's (1996) *'lexical organization'* and Henriksen's (1999) *'depth of knowledge'*. Accordingly, depth of vocabulary knowledge can be conceptualized in terms of how words are stored in the mental lexicon in an organized structure by means of processes of semantic mapping, collocation, synonymy, super-ordination, and co-ordination, which, in turn, are considered aspects of depth of knowledge in the literature of vocabulary assessment.

A similar way of conceptualizing depth of vocabulary knowledge is held by Meara (1996) who describes depth of vocabulary knowledge as a network of lexical items

greatly structured in the mental lexicon. Instead of depth, he favours lexical organisation. Put differently, how words are learned not how well words are well known or neither how many words are known when considering vocabulary breath. Lexical organization can be measured by means of assessing learners' ability to identify or give a large set of associations; however, depth conceived as lexical organization is considered as one measure of vocabulary breadth (Schmitt, 2014).

The above three dimensions indicate that knowing a word requires many aspects and these categories provide clear guidelines for vocabulary measurement but "each one is limited in the degree to which depth is measured, and overlap between the categories suggests that together they may still not provide an accurate assessment of depth" (Webb, 2013, p.3). In a similar vein, Nation (2001) further developed the above three aspects of depth in his famous table (Table 1.1, p.29) that sums up features included in mastering a lexical unit of language.

As illustrated in the table coming next, the three rudimentary constituents of form, meaning and use, as proposed in Nation's (2001, 2013) hierarchical table constitute deep word knowledge. These constituents, in their turn, split into three further sub-constituent aspects, and each aspect is also segmented into receptive and productive knowledge types (18 aspects). It is important to mention that the productive knowledge written in bold (in Table 1.1, p.29) form the backbone for the recent research. The bolded aspects of vocabulary knowledge that will be examined in the current test fit to the same textbook outcomes and content standards on which task instructions will be built.

Nation (1990) stipulates, in his table, that vocabulary knowledge is composed of active and passive vocabulary that correspond to productive and receptive vocabulary respectively. González-Fernández and Schmitt (2020) draw a border line between active and passive vocabulary stating that the former is named recall which refers to the learners' ability to retrieve word knowledge, and the latter is labelled recognition which is the learners' ability to perceive and select word knowledge. Possibly the best description of productive versus receptive vocabulary based on Nation's (2001) is set up by Schmitt (2014):

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	**R**	What does the word look like?
		P	**How is the word written or spelled?**
	Word parts	**R**	What parts are recognizable in this word?
		P	**What word parts are needed to express the meaning?**
Meaning	**Form and meaning**	**R**	What meaning does this word form signal?
		P	**What word form can be used to express this meaning?**
	Concept and referent	**R**	What is included in the concept?
		P	What items can the concept refer to?
	Associations	**R**	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	**Grammatical functions**	**R**	In what patterns does the word occur?
		P	**In what pattern must we use this word?**
	Collocations	**R**	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use (register, frequency...)	**R**	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

Table 1.1: What is Involved in Knowing a Word (From Nation, 2001, p. 27)

Note. In column3, R=receptive knowledge, P= productive knowledge.

A very widespread distinction following on from this approach is receptive versus productive mastery of an item (sometimes referred to as passive and active mastery, respectively). Receptive mastery entails being able to comprehend lexical items when

> listening or reading, while productive mastery entails being able to produce lexical items when speaking or writing. (p.919)

From the above extract, it follows that active vocabulary is associated with productive language skills of speaking and writing and passive vocabulary is related to receptive language skills of listening and reading. In other words, when a learner encounters a word form in listening or reading passages, s/he recognizes it and thus recalls its meaning, this is a clear case of receptive vocabulary knowledge. In productive vocabulary, however, a leaner is expected to use appropriate word forms corresponding to writing and speaking settings after recall and making choices about which word forms and meanings best convey the message appropriately.

As to the quality of word knowledge, Webb (2013) asserts that depth of vocabulary knowledge is identified by the extent to which Nation's (2001) eighteenth (18) aspects involved in knowing a word are present, and these in turn, would indicate whether a lexical item is fully mastered or not. These features involve both receptive and productive use ability. Qian (1999) has previously conceptualized depth dimension form a similar perspective assimilating syntactic, semantic, morphemic and graphemic characteristics of word knowledge.

The aforementioned conceptualizations indicate how vocabulary depth is a controversial and multifarious theoretical construct and this is what "makes it extremely difficult to know how to approach [it] from a theoretical perspective" (Schmitt, 2014, p. 915). However, the most elaborated descriptive framework of what is involved in deep word knowledge is that of Nation (Read, 2004; Deller et al., 2007; Gyllstad, 2013; Webb, 2013). In sum, knowing a word does not entail only knowing its form meaning connection rather it implies its semantic properties, its orthographic (spelling), phonological (pronunciation), morphological (word parts), syntactic (grammatical functions), collocational (word associations) and pragmatic (register) characteristics.

As the researcher adheres to Nation's multidimensional framework of depth of vocabulary knowledge, it is worthy to consider its aspects and how to test them.

1.4. Nation’s Aspects of Vocabulary Knowledge and How to Test Each Aspect

Receptive and productive aspects of vocabulary knowledge stretching from Nation’s hierarchy seem worth a concern. Passive and productive notions all-encompassing vocabulary knowledge are further described along the following lines in relation to the testing procedures that are relevant to these two vocabularies.

1.4.1. Vocabulary Knowledge: Receptive Vs Productive Aspects

Vocabulary tests often reflect the designer's definition of word knowledge, focusing on individual words or formulaic expressions. Nation (2013) posits that lexical items derive meaning only when contextualized within interrelated systems and levels of a language. Assessing vocabulary knowledge involves considering various linguistic features that determine the extent of understanding.

At a general level, Nation (2013) identifies form, meaning, and use as essential components of knowing a word. Using the word 'underdeveloped' as an example, he illustrates the continuum between receptive and productive knowledge. The receptive knowledge of ‘underdeveloped’ involves four major aspects: recognizing the word's spoken and written forms, understanding its meaning in context, identifying related words and typical collocations, and finally, recognizing its appropriateness and frequency of use. The productive knowledge of this word, on the other hand, entails learners’ ability to pronounce it correctly, spell it accurately, use it appropriately in speaking and writing, employ the words’ synonyms and antonyms, and above all, using the word in various contexts and recognizing its degree of formality and informality.

This distinction underscores the complexity of vocabulary knowledge, highlighting the need for comprehensive assessment and effective teaching strategies.

The above aspects dichotomously scoped the two dimensions of receptive and productive knowledge. They both represent levels of deep word knowledge for testing. The productive dimension serves the current research purposes especially production and use of words in context, with correct spelling, correct word forms (derivations and inflections).

1.4. 2. Aspects of Word Knowledge for Testing and How to Test Them

From Nation's (2001, 2013) hierarchical table (Table 1.2), which is included below, it implies that deep word knowledge comprises three basic components; form,

Form	Spoken	**R**	Can the learner recognize the spoken form of the word?
		P	Can the learner pronounce the word correctly?
	Written	**R**	Can the learner recognize the written form of the word?
		P	Can the learner spell and write the word?
	Word parts	**R**	Can the learner recognize known parts in the word?
		P	Can the learner produce appropriate inflected and derived forms of the word?
Meaning	Form and meaning	**R**	Can the learner recall the appropriate meaning for this word form?
		P	Can the learner produce the appropriate word form to express this meaning?
	Concept and referents	**R**	Can the learner understand a range of uses of the word and its central concept?
		P	Can the learner use the word to refer to a range of items?
	Associations	**R**	Can the learner produce common associations for this word?
		P	Can the learner recall this word when presented with related ideas?
Use	Grammatical functions	**R**	Can the learner recognize correct uses of the word in context?
		P	Can the learner use this word in the correct grammatical patterns?
	Collocations	**R**	Can the learner recognize appropriate collocations?
		P	Can the learner produce the word with appropriate collocations?
	Constraints on use (register, frequency …	**R**	Can the learner tell if the word is common, formal, infrequent, etc?
		P	Can the learner use the word at appropriate times?

Table 1.2: Aspects of Word Knowledge for Testing, Nation (2013, p. 538*)*

meaning and use; and each of which is composed of three other subcomponents aspects, and each aspect can be further divided into receptive and productive knowledge categories. This suggests that the targeted productive word knowledge in this research comprises basically knowing its meaning, form and use. These aspects are not exclusive since they cannot be covered in one test battery. Some of these levels might be selected as to reflect assessment purpose, the core of the course content and the materials being covered.

When vocabulary knowledge is assessed not only word meaning aspect is to be considered but, instead, manifold aspects can be incorporated particularly when depth and not breadth is being targeted. Nation (2013) clearly indicates how each of these aspects can be evaluated by means of 18 test item types that are closely connected with Table 5. These are summarized along the following lines:

The spoken form of a word or sentence can be tested using dictation, translation to L1, (to test learners' recognition of the spoken form) and reading aloud or cued oral recall to check whether learners can correctly pronounce the word. The second aspect, written form, uses measures like pronouncing written words in order to test written form/word recognition, and dictation which is a productive vocabulary test is meant to test learners' ability to spell or write words. The third aspect named word parts, is tested by breaking the word into parts (morphemes or affixes), or selecting or giving meanings of the morphemes, or produce an affixed form of a familiar word. The fourth aspect called form and meaning is tested via translating into L1 or L2 or selecting the appropriate picture. The fifth aspect, concept and referents, is measured by means of translation (e.g., to find out equivalents into L1 or to select proper words to translate into L1). The sixth aspect known as associations is tested either by selecting word's associations or adding associates to an existing list. The seventh aspect, grammatical functions, is assessed via sentence correctness recognition or applying knowledge of words at the level of sentence context. As for collocations, they are assessed by means of recognizing sentences as being correct or incorrect and producing collocations. The last aspect related to constraints and word frequency is assessed by asking leaners about

their knowledge of word use (E, g, which of these words represent UK use? and register (e, g, what is the formal word for X?).

Now that the aspects of receptive and productive vocabulary and procedures on how to test them have been delineated, the subsequent sections will tackle different types of test formats used in the literature of vocabulary assessment and research.

1.5. Types of Vocabulary Tests

In this section, we will review the way vocabulary size and depth of knowledge have been operationalized and hence tested in previous research up to 2022. Most vocabulary tests turn around measuring meaning recognition, word use, meaning production, or word associations. Others, however, trace learner's vocabulary development along a continuum of knowledge.

1.5.1. Vocabulary Size Tests

Vocabulary size tests (VSTs), also are referred to as vocabulary breadth tests in the literature of vocabulary assessment, tend to calculate the size of learners' vocabularies, i.e. how many words and word families learners know (Nation, 2001). They are severely criticized for a non-profound representation of lexical items (Schmitt, 2000). They are used to just focus on one level of vocabulary knowledge (word recognition). One advantage of VSTs is its ability to incorporate a large sample of items this, in turn, makes it possible to assume that test results represent to a certain extent test takers' vocabulary repertoire. In this regard, Read (2000) suggests that, unlike depth tests, size tests (STs) sound to be superficial, but still have good quality to depict learner's total vocabulary repertoire. However, this test still approaches vocabulary discreetly disregarding context. One famous example of STs which is known as *Vocabulary Levels Test* (VLT) has been is developed by Schmitt, Schmitt, and Clapham (2001).

These authors' (2001) VLT, a measure of meaning recognition, focuses on high frequency vocabulary. The version contains five levels, as described by Schmitt et al. (2001), the 2,000 word level of high frequency vocabulary; the 3,000 word level of low frequency; the 5,000 word level of low frequency words; the AWL high frequency word

for academic purposes; and the 10,000 word level low frequency. Each level tests 30 words as each level is composed of ten blocks of items., Schmitt et al., (2001, p. 82) provide the following format:

Excerpt from Appendix II from student instruction sheet for the VLT (Block 2).

You must choose the right word to go with each meaning. Write the number of that word next to its meaning.

copy

event - end or highest point

motor - this moves a car

pity - thing made to be like another

Profit

tip

The VLT version that is mostly used in the literature of VS has plausibly demonstrated a valid measure for arguably providing a high estimate of test takers' vocabulary knowledge at diverse levels frequency (Schmitt et al., 2001).

Form this test format, it sounds feasible to say that it tests receptive knowledge of words. That is, learners have just to select appropriate meanings from lists of words containing distracters. Productive Vocabulary Levels Tests (PVLTs) are used for diagnostic purposes compared to VLTs that aim to have an estimate of high and low frequency vocabulary known by a learner. PVLTs are characterized by some degree of productivity elicited by means of word completion tasks (as in Laufer's and Nation 1999).

In a review of research conducted on vocabulary breadth, Anderson and Freebody (1981) questioned the confidentiality (credibility) of these types of tests, as they have resulted in such distinct degrees of estimates of words known by a particular group of learners or children and even among educated adults. Concurrently, Daller et al. (2007) assume that it is unusual to ask someone about the number of words s/he knows in a language because the question itself does not have a clear answer, in simpler terms, a

word itself does not have accurate definition because of its complex nature (see threats of vocabulary validity section, 2.7, p. 145). Definitely, "Knowing a word involves more than knowing a word's definition" (Johnson & Pearson, 1984; Nagy & Scott, 2000, as cited in Stahl, Bravo, 2010, p. 567-568). Besides, the type of knowledge elicited in VSTs is principally "basic form-meaning mapping" Gyllstad (2013, p.14).

In short, despite their advantages in providing better estimates of words known by learners, and despite the fact that the degrees of reliability and practicality that yield in are rather high, VSTs focus on the quantitative nature of participants' word knowledge at the expense of other aspects. This made them unsatisfactory as more information about the quality of knowledge is rather needed to get a better, and perhaps deeper, understanding of students' lexical ability, be it productive word use in contextualized language setting language. Given what has been said, "there has been an increasing awareness that there is much more to knowing a word than just learning its meaning and form" (Schmitt & Meara, 1997, p.18). This recommended trend is extensively discussed below.

1.5.2. Depth Tests

Alternatively, depth tests (DTs) are applied, against and, as a reaction to VSTs. In DTs each lexical item is treated from multiple dimensions of knowledge. Adversely, the number of items in breadth tests is restricted and, therefore, these tests cannot claim to represent learners total vocabulary threshold. According to Schmitt (2000), DTs aim to measure how well (the quality of knowledge) learners' vocabularies are recognized. On the quality and volume of word knowledge that a learner might have, Stahl (2018) provides us with the following figure (Figure 1.2) that displays the two aspects of voca-

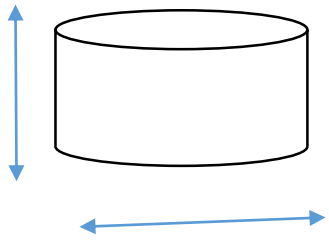


Figure 1.2: The Challenge of Vocabulary Assessment (Stahl, 2018, P.5)

vocabulary knowledge tests. Stahl (2018) proposes a number of assessment methods namely, assessment breadth (volume of words), assessment of depth via reading, listening, speaking/oral usage of words, writing/written usage of words (sentences/passages using the words).

Our main concern is not to gain an estimate of vocabulary that learners know. VSTs test only a single aspect of knowledge, which is meaning recognition and recall, neglecting other aspects representing reciprocal basis of deep knowledge tested. The former test format does not seem useful in testing deep vocabulary knowledge. Depth is measured by various test formats and approached diversely being a multifaceted construct containing multiple aspects involved in mastering a lexical item.

1.5.2.1. Approaches to Measuring Depth of Vocabulary Knowledge

Depth of vocabulary knowledge measures developed in search of assessing the quality of vocabulary knowledge are relatively few (Mukarto, 2005; Schmitt, 2014; Ebrahimi, 2017; Edmonds et al., 2022) especially when compared to VSTs. One reason for that is related to the fact that breadth tests are easy to administer and can embrace extensively large quantity of words in a rather short period of time as they merely need a single response format associated with every lexical unit (Read, 1993).

Read (2000) distinguishes two approaches to measuring deep knowledge of target words. One is named the developmental approach and the other is called the dimensional or componential approach. Recently, however, Yanagisawa and Webb (2020) add one more approach to measuring depth, known as the lexical network approach. In the first approach developmental scales are utilized to identify the phases of mastery of a lexical item and, as such, it embodies the incremental nature of vocabulary learning (the scales range from no knowledge labelled as 0 to complete mastery labelled as 5). For Read (1993), it is because of the fact that vocabulary knowledge is frequently a gradual process (even native speakers of a language have an imprecise knowledge of meaning of words they know) that many researchers, when measuring vocabulary knowledge, devised different scales to represent different levels of word knowledge. Adversely, in the dimensional approach, researchers measure the

degree to which the multiple dimensions or constituents of vocabulary knowledge have been mastered. Overall, researchers' adherence to these approaches depends on how depth of vocabulary knowledge is devised; whether is conceived of as degrees of knowledge, as lexical mental connections, or as multiple sub-knowledge aspects.

In the literature of vocabulary assessment, a number of methods have been well-established to assess how well words are known. To name but few, the *Vocabulary Knowledge Scale* (Wesche & Paribakht, 1996) constructed in the developmental approach paradigm, the *Word Association Test* (Read, 1993, 1998) and the lexical network test or *V-Links* (Meara, 1996; 2009) associated with the lexical network approach, and the *Interview* (Schmitt, 1998) produced in the components approach. In the upcoming section we will review the measures of depth of vocabulary knowledge that have been employed in previous research.

1.5.2.1.1. Developmental Approach

Unlike the dimensional approach, the developmental approach uses only one assessment instrument, that of *Vocabulary Knowledge Scale* (VKS), also known as self-report of degrees of knowledge (Read, 2004). It is a test battery that was first produced by Wesche and Paribakht (1996). It can be used as a measure of productive as well as receptive vocabulary depth since it contains two different levels of scales; one involves learners to use particular words in a sentence and others require no creative use of the word, thus, tapping just the passive knowledge. The VKS has been described as a generic measure in that it can be applied to assess mastery of any set of words (Wesche & Paribakht, 1996; Read, 2000). It uses five scales; four are rather similar to Dale's (1965) in addition to a fifth stage under the statement 'I can use this word in sentence'. From this scale it can be inferred that besides recognizing a word's meaning and its appropriate place in context, a learner can further use it in sentence context. Initially the scales chart four levels of word mastery especially developed by Dale who states that the incremental nature of vocabulary knowledge is due to four stages of vocabulary growth. He posited the existence of, at least, four incremental stages of word knowledge upon which Wesche and Paribakht (1996) elaborated their most sophisticated five point self- report elicitation scale as reflected in the research literature:

1. I don’t remember having seen this word before.
2. I have seen this word before, but I don’t know what it means.
3. I have seen this word before, and I think it means _____. (synonym or translation)
4. I know this word. It means _____. (synonym or translation)
5. I can use the word in a sentence. (p. 30)

The above stages have been adopted as a measure in vocabulary assessment, especially assessing the quality of vocabulary knowledge. It is highly important to point out that the major difference between Dale’s and Wesche and Paribakht’s scales is that the former designed his scale for L1 learners whereas the latter developed it for SL users. The test was designed in the ESL vocabulary development assessment endeavor. It was constructed basically to indicate how receptive knowledge of some words would change to an initial productive knowledge over time. This growth in vocabulary learning is a result of vocabulary instruction through reading and other activities.

Despite the fact that it has sustained great attentiveness as a proposal in vocabulary assessment (Schmitt, 1998), the VKS has received severe criticism for being time-consuming research tool as it tests only a limited number of words. It is time consuming because it targets both spoken and written explanations of words from the part of learners (Read, 1993). Another limitation that can be addressed to this test battery is that it falls upon discrete point testing that measures words out of their larger communicative context. Eventhough one component of the scale measures word use but emphasizes context at the sentence level.

1.5.2.1.2. Lexical Network Approach

The lexical network approach endeavors to investigate how words are constructed or stored in the mental lexicon; that is how words are linked to each other in someone’s mind (Read, 1993; Meara, 1996; Henriksen, 1999). It deals with depth of vocabulary knowledge as organized lexical sets (or networks) in L2 learners’ mental lexicon. This approach had its genesis in Read’s (2004) conception of deep word knowledge, that of network knowledge (the third type, see section 1.3, p. 24).

Two very famous successful measures adopting the lexical approach have been

taken up by previous research: *The Word Association Test (WAT)* developed by Read (1993) and the *Vlinks (lexical networks)* produced by Meara (1996). The WAT receptively measures how well specific target words are well known. Read claims that examinees who are able to track the associations seem to have deeper lexical knowledge and those who are not their knowledge is fairly superficial. In this typical lexical network test, learners are asked to recognize or supplement a number of associations. In this regard, Schmitt and Meara (1997) define word associations as "the links that connect or relate words in some manner in a person's mind. A common way of eliciting them is to have a tester give a prompt word and have the subject say the first word that comes to mind" (p. 19). Word association tasks require a number of features that are explained as under.

In word association tasks, learners are required to identify lexical or semantic networks related to prompt words which are presented with a set of other words (often are 4 associates), some of which are lexically or semantically related to the stimulus and some others do not and act as distractors which either resemble stimulus words orthographically or morphologically. The test attempts to elicit three types of semantic networks existing between the words; whether the lexical items are connected syntagmatically (collocations); paradigmatically (synonyms); or analytically (the associate partially represents the meaning of the stimulus word). Let us tackle one example from the WAT format:

Example: *denominator* (stimulus word)
common develop divide eloquent fraction mathematics species western
(Read, 1993, p. 366). Here the test takers are instructed to pick out four words that are associates; either lexically or semantically related to the stimulus word.

All-inclusively, the test examines how deeply learners know 40 target words (nouns, verbs and adjectives) by selecting four words, out of eight words, associated with the stimulus word. This testing procedure arguably proved to be a reliable measure of depth of vocabulary knowledge, however it is merely a discrete measure as it approaches lexical items in isolation, and is definitely receptive in that learners are

required to select correct answers from odd words instead of creating or supplying their own associates.

Vocabulary/Lexical Networks:

It has been mentioned in the previous section that there are two distinct perspectives to depth of vocabulary knowledge, one of which conceptualizes depth as a set of sub-knowledge components and the other, however, prefers the term organization instead. The latter approach typically refers to Meara (1996) who has introduced his idea of organization to mean that strong structured mental networks of lexical items. His approach to vocabulary testing is developed out of, and as a reaction to, VS and DT tests. The assumption is, testing knowledge of a large number of words needs extensively developed measures for testing individual words and even including all aspects of word knowledge to depict how well 50 words, for instance, are known makes the process of assessment impossible and even unreliable as it requires designing dozen of subtests (600 item test). An alternative approach to these is to produce a single test that can assess depth meaningfully. He, thus, designed his most well-known V_Links, a test battery to measure lexical organization in 2009. It is a 20 items test that are randomly selected from the basic 1000 words in English core vocabulary and each of which comprises a selection of 10 target items. It is instantiated in a software program in which subjects are asked to recognize any sort of association pairs and show how far the association is strong on a four point scale. According to Nation (2014), this research endeavor is marginalized and still sophisticated and he emphasizes that large scale research is recommended to further investigate whether this approach can be optimized and widely realized.

Though it proved to be an effective measure of lexical organization as it extensively tests large sets of words (a total of 200 words) in a relatively short period of time (30 minutes), it neglected the larger communicative aspect of language, i.e. testing word use in context is rather ignored.

1.5.2.1.3. Components/Dimensions Approach

This approach conceives of vocabulary knowledge as comprising a set of levels of knowledge. It adheres to Cronbach's (1942), Richards' (1976), Henriksen's (1999), Read's (2004) and Nation (2001, 2013) operationalization of depth of vocabulary

knowledge. However, Nation's framework of what is involved in knowing a word arguably proved be the best comprehensive basis for modelling and assessing word knowledge (Webb, 2013; Yanagisawa & Webb, 2020). It is obvious that the multidimensional nature of vocabulary made it difficult and impossible to design test batteries integrating the multiple aspects of depth. Subsequently, a number of assessment methods have been established in the literature of vocabulary assessment. These involve interviews (e.g., Schmitt, 1998), *idioms method*, *Context-Dependent Collocation* (Milton, 2009), and procedures for *Meaningful Sentences* (Sharakhimov & Nurmukhamedov, 2021), and *Depth Tests* (Read & Dang, 2022).

In a longitudinal study, Schmitt (1998) produced an interview procedure, summed up along Table 1.3, to assess the quality and growth of learners' vocabulary knowledge. The study aim was to track the acquisition of 11 words of three proficient adult university FL learners coming from different backgrounds over a course of a year. Four types of word knowledge namely spelling, associations, grammatical information, and meaning have been measured after test takers were exposed to the words during the course of study. Target words were selected from UWL (University Word List: academic vocabulary) they were all polysemous words having more than three senses to raise learners' awareness of various senses. Schmitt designed a test of written form (spelling) and association measurement procedure that required learners to provide 3 responses corresponding to target word stimulus.

	Interview procedure	
1	How do you spell _____?	To test written form
2	Please give the first 3 words you think of when you hear the word _____.	Association measurement procedure
3	What word class (part-of-speech) is _____? Is there a (noun, verb, adjective, adverb) form? If so, what is it?	Grammatical knowledge elicitation
4	Explain any meaning senses you know for the target word? (give definitions, give examples, use the word in sentences, draw sketches or diagrams, use gestures, etc.)	Knowledge of meaning senses

Table 1.3: Test Sheet for Interview Procedure (Schmitt, 1998)

Eventhough it enables vocabulary researchers to tap extended information needed to measure depth of lexical knowledge, this interview instrument is criticized for some pitfalls. It is time consuming and this, in turn, limits the number of target words as well as research participants. Furthermore, it limits the scope of contextualization to the sentence level instead of large situational communicative context of language use.

Milton (2009) proposes another measure to assess test takers' knowledge of fixed English idioms. In this test format, known as *idiom knowledge*, learners were given 20 sampled idioms and were instructed to supply a missing word for each gap in the target idioms. Idioms are essential component in depth of vocabulary knowledge; nevertheless, Milton questioned their usefulness for testing L2 knowledge, as they are likely to be rarely used in the actual native language usage because they are highly infrequent in normal speech.

Sharakhimov and Nurmukhamedov (2021) designed two insightful tests. These are *Context-Dependent Collocation* and *Meaningful Sentences* procedures to assess learners' quality of word knowledge. The first procedure endeavors to test examinees' ability to produce collocations by means of asking them to produce two or more word combinations of words that frequently co-occur. Colocations are questionable to vocabulary acquisition because, as Webb and Sasao (2013, p. 269) state, learners "often know words but are unable to use them effectively because they do not know their collocates". The second assessment procedure measures productive vocabulary use whereby respondents are required to produce meaningful utterances using target words. Sharakhimov and Nurmukhamedov (2021) recommend the second procedure to be part of listening and speaking activities (conversations). Nevertheless, given that the Context-Dependent Collocation and Meaningful Sentences procedures focus on using words to construct word combinations like phrases, and to produce meaningful sentences respectively prevent depth of vocabulary from integrative, communicative, situational and contextualized assessment.

More recently Read and Dang (2022) suggest a *Depth Test* that endeavors specifically to measure the quality of students' knowledge of high-frequency academic vocabulary in English. It assesses knowledge of synonyms, collocations, and word parts of known words. The test consists of three parts: Part A is intended to measure if leaners

can match the target words with a synonym or short definition. It uses a simple selected-response format, and Yes/No format. It introduces two synonyms and two distractors for each target word with a total of 30 items. Part B, devoted to testing collocations, involves 30 whole phrases and short sentences to assess knowledge of collocations using yes/no and not sure options in order to indicate correctness of word combinations. Part C requires examinees to find out the correct word form using both inflected and derived affixes. There were 20 target words in this part, yielding to a total of 60 items. For individual words, there were one or more correct word forms. Though proved to be more reliable and insightful, these three depth tests could be questioned in terms of context, which plays a significant role in getting a clear image of students' ability to use words to communicate effectively via writing or speaking, because their overall purpose is directed towards selective response rather than constructed response performance.

Everything being equal, the aforementioned depth tests though seem to be discerning, well established and most influential, they do not bother about measuring vocabulary in embedded comprehensive context-dependence assessment methods. They neglect the communicative endeavor of language and they do not provide learners with authentic communicative situations whereby their production and word use ability will be elicited. It is worth mentioning that the proliferation of WAT format has arguably proven to be the most useful method in the research paradigm, especially being applied to a large extent by many researchers (e.g. Stahl, 2018) due to its points of strength. Even so, the WAT will not be adhered to the current research paradigm as it neglects the communicative context of tested words.

The issues reviewed so far point to some inaccuracies that can be demonstrated through various depth of vocabulary assessment methods. For the purpose of the present study, reviewing the literature on productive vocabulary tetsing is needed in order to develop effective assessment of depth of productive vocabulary knowledge.

1.5.2.2. Tests of Productive Knowledge

The vocabulary learning process has long been most importantly investigated along receptive and productive continuum particularly learners' vocabulary knowledge in use (Zhong, 2018). The term *productive* for Daller et al. (2007) means that "the subject has to write-produce- the word rather than recognize it" (p.121). In the literature

of vocabulary assessment, two types of productive vocabulary knowledge tests have been developed. These consist of controlled productive knowledge tests and free productive knowledge tests. The former is measured by means of Productive Vocabulary Levels Test (PVLT) as developed by Laufer and Nation (1999, *for more information on test format visit the Compleat Lexical Tutor*: https://www.lextutor.ca.). Laufer and Nation's (1999) version of PVLT is made up of five parts, each part tests productive knowledge of 18 target words (with a total of 90 target words) to determine the frequency level and/or the academic level of participants. The frequency levels displayed in the test version are the 2,000 level, the 3,000 level, the 5,000 level, and the 10,000 level, (the frequencies contain the most frequent words used in the target language). The academic level samples words from Coxhead's (2000) Academic Word List (AWL), based on word families that are mostly frequent in academic texts.

The controlled vocabulary measure is composed of items selected from five frequency levels deploying a completion item-type format. Participants are required to complete the underlined words. The example has been provided for test takers: He was riding a bicycle (we extracted one example per each word frequency list from Laufer and Nation (1999, p.46-47).

The 2000-word level

The pirates buried the trea___ on a desert island.

The 3000-word level

France was proc___ a republic in the 18 th century.

The 5000-word level

The angry crowed sho__ the prisoner as he was leaving the court.

The University Word List level

According to the communist doc__ , workers should rule the world.

The 10 000-word level

The baby is wet. Her dia__ needs changing.

The above extract implies that productive vocabulary tests measure "the ability to

use a word (…), whether in an unconstrained context such as a sentence writing task, or in a constrained context such as a fill in task where a sentence context is provided and the missing target word has to be supplied" (Laufer & Nation 1999, p. 37). For these researchers this test, which is meant to distinguish between high and low proficiency learners, proved to be practical, easy to administer, easy to score and interpret, reliable and valid. Nevertheless, the current study probes to concentrate on the first type (productive vocabulary knowledge (PVK)) but rather in an extended context. Focus should not be put on context at sentence level but extends to context at the paragraph or written discourse level.

Some researchers adhere to another measure in the testing of productive use of vocabulary, Zhong (2016), for example, investigated the relationship between receptive knowledge of meaning, form, word class, collocation and association and productive vocabulary knowledge, notably the controlled productive word use in sentence writing, with a multi-task approach. His participants were 620 Year 8 EFL learners. Twenty-six target words were sampled from participants' textbook in both the pre-and post-test design. He used distinctive research procedures: Form recognition, Meaning comprehension, Word class knowledge, Association, Sentence writing. A revisited version of the VKS (Wesche & Paribakht 1996), which presents test items in a checklist format, the aim of which is to check depth of word knowledge receptively and productively (meaning comprehension task). A Recognition Task was applied to test receptive word knowledge in a multiple-choice word form (spelling) where test takers were required to choose the correct word from three distractors to check students' recognition. A fill-in-the-table task was employed to measure students' knowledge of word classes. To capture knowledge of association and collocation, the Word Association Test (WAT) was adapted from Read (1998). Students were asked to select four out of eight words whose meanings associate or collocate with the target word. Learners' PVK was elicited by means of sentence writing were they were asked to write from two to three sentences, one of them should encompass the targeted word. Research findings indicated that productive word use is associated with receptive form and meaning knowledge in prompted sentence writing.

The research findings indicated how word class, association and collocation are significant factors in word use. They also suggest the significance of from and meaning in the teaching and learning of vocabulary and provide quantitative evidence that show why measuring meaning and form only is cost-effective in vocabulary testing. Findings also suggest that other lexical aspects cannot be explored if part of the research objective is testing productive word use. The results further indicated that collocation and association, as aspects of depth of vocabulary knowledge, resulted in increased variation in productive word use, and when exceedingly mastered these aspects would yield in more confidence once used in sentence production.

This controlled productive vocabulary test format adheres to the dimensional approach to testing deep word knowledge as it emphasizes word form, meaning and use. However, the latter aspect seems critical especially in terms of contextualization. Like PVLT, vocabulary in the above study is measured by means of sentence writing with target words. Vocabulary assessment is restricted to sentence level context rather than contextualized in larger contexts or discourse, a common feature in communicative and task-based settings.

A review of the literature on the free productive vocabulary knowledge methods reveals that it has been measured predominantly through lexical richness and association tasks. One famous method used for measuring lexical proficiency by means of calculating lexical richness is the Lexical Frequency Profile (LFP) developed by Laufer and Nation in 1995. The aim of which is to quantify the extent to which L2 examinees are using varied and large vocabulary (East, 2004). One more measure of free productive vocabulary knowledge is Meara and Fitzpatrick's (2000) free word association task named Lex 30 used especially to measure depth of vocabulary knowledge (Agdam & Sadeghi (2014).

Test batteries developed on measuring lexical richness and lexical sophistication and diversity have not been discussed so far in this section because they do not fit the research endeavor that seeks to assess the quality of learners' word knowledge. How well learners can recognize meanings of target words, and their ability to use word formation strategies to create new word parts and thus use them to sole situation

problem scenarios. The aim, therefore, is neither to elicit lexical richness nor sophistication and diversity.

The different types of tests explored so far, whether set up for receptive or productive knowledge, or even for measuring depth of vocabulary knowledge reveal a number of shortcomings. These are briefly examined.

- Learners' receptive vocabulary knowledge is favored to a large extent at the expense of productive vocabulary knowledge (except for LFP and WAT measuring meaning production at sentence level context);

- An overemphasis on decontextualized design measures in favor of partial declarative knowledge "form-meaning connections" (e, g. MCQ standardized testing). Measures of this kind are said to be sensitive as described by Nation (2013);

- A variety of test batteries tap different types of learners' knowledge of lexical items. This illustrates the discrepancies and controversies involved in knowing a word and what components or levels of these words to measure. This feature made these tests with different levels of sensitivity (Nation, 2005);

- More stress is put on the assessment of vocabulary as part of a larger construct; as part of writing proficiency as illustrated in measures of vocabulary richness and sophistication;

- Decisions made on lexical selection is based on well-known frequency lists and word counts but not on textbook or classroom vocabulary, that is, what is taught and learned during a course of instruction. Achievement and diagnostic tests tend to assess what words have been learned after a course, and this sort of vocabulary is less approached or is often limited; and

- Ultimately, the tests discussed so far are tests of declarative rather than procedural knowledge. Nation (2013) defines the former as the conscious deliberate retention of vocabulary concerned mainly with form-meaning connections, in brief, it is all about that knowledge that learners can declare. The second type, however, refers to learners' receptive and productive knowledge or simply that ability to use words while receiving and conveying messages. In this realm, vocabulary in use or vocabulary fluency is far

less approached within larger discourse, and if so, it is treated as part of larger construct, and when contextualized the context is limited to sentence context.

Overall, each test of vocabulary has its strengths and pitfalls. The researcher gained a sense of inspiration along the limitations of assessment procedures discussed earlier. The researcher suggests more focus on assessing learners' ability to recognize meanings of target words, construct appropriate word parts with their corresponding inflections and appropriate spelling, and use those words in large communicative contexts containing situations and prompts. In a controlled depth of productive vocabulary knowledge test, learners are required to use target words to write short constructed responses (paragraphs) respecting the contextual themes and grammatical boundaries of those words put within a set of organized linguistic choices. Target words must have been familiar, that is covered during a course of instruction within reading and listening activities when considering incidental learning, or sections devoted to teaching vocabulary in intentional vocabulary instruction as to cater for achievement testing.

Learners' lexical competence, therefore, will be elicited by means of designing a set of communicative, complex, authentic tasks where leaners are asked to produce pieces of written discourses where they use target words to solve new problems set up in situation problems scenarios. As such, context plays an important role in assessing learners' vocabulary knowledge, a core issue discussed in the next section.

1.6. Vocabulary Assessment in Context

After a course of instruction, teachers, in their everyday classroom practices, often make use of a variety of methods to assess students' vocabulary knowledge and development. Learners might be instructed to define a word (provide a short dictionary definition), select the correct word from a list of responses (case of multiple choice questions containing distracters), fill in the blank, identify the opposite or synonym when considering sense relations, matching words, …etc. Eventhough, these ways might give teachers a clear image about their learners' progress, they are not, it seems, the best measures as they emphasize either word recognition or meaning recall but, definitely, not meaning construction (e.g., construct the antonym and use it to write

correct/meaningful sentence). Put differently, “These modes of vocabulary assessment are shallow metrics of possible word knowledge” (Stahl & Bravo, 2010, p. 566). Overall, in order for the test to be more effective in testing language ability, it is suggested to engage learners in activities where they display their ability to use the target words in context to communicate in every day writings and conversations.

Among the factors that might influence FL production is contextualization, the communicative aspect highlighted in this study. When considering the nature of language as communication, every single word should have its contextual usage because “[t]eaching vocabulary requires situating the word within a system of ideas to be developed” (Stahl & Nagy, 2006, as cited in Stahl, Bravo, 2010, p.566). In this very specific context, a full mastery of a word encompasses different aspects including, among other things, semantic and syntactic categories, and most importantly word use in context. Testing learners’ ability to correctly use words in context would elicit their word knowledge and mastery of word meanings and forms and, hence, lexical sentence boundaries. One measure to achieve this is to gauge learners to construct simple sentences or short paragraphs subsuming their knowledge of synonymy, antonymy, polysemy, word classes, or dictionary meanings, etc. On the significance of context in EFL, Schmitt (2014) posited that,

> Studying the words in isolation without contextual elaboration limits the students to learning only something about the word form, something about the meaning, and some linkage between the form and meaning. To the extent that the students are able to retain some of these form–meaning connections (and this may be difficult without consolidation), then the students may have a relatively larger vocabulary size, but would know relatively little about these words, and probably would not be able to use them to any great degree. (p. 915)

Following this line of thought, it becomes obvious that teaching words in discrete points yields in grasping form-meaning mapping, this in turn, deprives learners form enhancing their productive vocabulary use. According to Schmitt (2014), a word’s collocations, register constraints, and frequency require distinct contextual use in language.

The issue of developing measures whereby vocabulary depth should be tested in

context was previously posed by Read (2004) in his *"Plumbing the depths: How should the construct of vocabulary knowledge be defined",* an article written in an attempt to conceptualize depth of vocabulary knowledge and formulate accurate measures for its assessment. Read (2004) comments on Nagy's and Scots (2000, p. 273) quote who conceptualized depth of word knowledge as being purely procedural and writes:

> Ultimately the question is not what learners know about a word but what they can do with it: being able to pronounce it, recognize it in connected speech and writing, and use it fluently in their own production. Thus, measures of declarative knowledge need to be complemented by tests of vocabulary in use in order to obtain a full picture of the learners' lexical competence. (p.224)

In line with the above quote, Read (2000) already posited that vocabulary assessment should be addressed from two different perspectives: Context dependent and context independent approaches. The former favors word use in context whereas the second emphasizes assigning vocabulary as a discrete point in a semantic field regardless of its context (For more knowledge see Table 1.4, page 59). Some insights in this table have inspired us to choose an appropriate test format to meet the research purposes among them is word use in context. But before making any decisions about test format it is important to discuss factors affecting test modelling and design.

1.7. Fundamental Issues in Modeling and Assessing Vocabulary Knowledge

When designing any vocabulary test, a range of factors intrude and affect the validity of a vocabulary measure especially when decisions about designing test format are being made. Nation and Read (1986), Nation (2005, 2007), and Daller et al. (2007) draw our attention to six factors that represent potential threats to vocabulary assessment validity. These consists of lexical item choice, test taker attitudes about testing, the nature of vocabulary itself, researcher's view of what aspects included in knowing a word, and vocabulary in context and use. Test construction, therefore, will be complex and too misleading when all these aspects are considered. These tricky and challenging factors and the way they might influence modelling[3] vocabulary assessment (test format, and its results henceforth) are further described below:

-***Vocabulary choice:*** It is very troublesome to select what words to test

eventhough many word counts are ready made and accessible for all researchers. Word selection for testing is a challenging obstacle for assessment stakeholders despite the appropriateness of frequency data developed by corpus linguistics.

-Test takers attitudes about the subject and individual variability: In a testing situation, students sometimes tend to be careless about their tests and their answers would be performed with no attention, especially with multiple choice test formats. Students might hold negative attitudes towards testing due to lack of training in test taking strategies and having bad test experiences. Milton (2008) picks up the paradox that test validity is based on the assumption that learners will behave in a reasonable consistent manner, yet as we all know, learners in a test situation do not necessarily do so; they may be unmotivated and give up, or they may take a strategic approach using guesswork. Such learner variability potentially comprises test validity. Milton also assumes that there is a predictable relationship between word frequency and acquisition. As for individual variability, Milton states that in the learning of high frequency vocabulary, high frequency vocabulary is learned first (learning variation).

-What is a word: The complex nature of vocabulary makes it very difficult and controversial to define exactly what a word is. It is correspondingly perplexing as to which words to include in a vocabulary test and the same does hold true for what component parts to focus on, i.e., the unit of counting used might be based on word families/lemmas; with base word/head word and its inflections. The basic question to be raised here is whether to consider multi-word units a unit of counting or to consider inflected words, or head-words, function words, content words, etc. If a researcher samples all these types of vocabularies considering all these forms as distinct words would increase an estimate of vocabulary size. One approach to solving this issue is proposed by Read and Nation (1986) who assume that the issue is no that difficult, because “the lexicon can be classified into lemmas (or word families), which can be represented for testing purposes by a base word. Thus, we assume that, if one knows the base word, little if any additional learning is required in order to understand its various inflectional and derived forms” (p. 6). According to these authors, once the issue of sampling has been solved, the types of knowledge is another doubtful issue to

consider.

-***Multiple aspects of word knowledge:*** Each target word should be tested in two or three levels (Read, 2000; Nation, 2007). This idea conditions the use of multiple measures reflecting the multidimensional nature of vocabulary knowledge. For example, to test the quality of vocabulary knowledge for L1 and L2 test takers, there is a need to consider measures of meaning, spelling and word parts, and measures that stress on actual language use to have a full image of how well-words are mastered. Based on this insight, this research will consider three major aspects in testing word knowledge, namely word meaning, word forms/parts and word use.

The conclusion drawn from these factors call for the need to do further research on systematic vocabulary knowledge measures or formats that are often hampered by lack of approved world-wide unified versions set up forward for testing. So far, this section described the factors influencing test content and construct validity. The next section is concerned with test design and development.

1.8. Vocabulary Test Development and Conceptualization

The complexities involved in assessing vocabulary depth, particularly productive depth, pose significant challenges. Defining the specific aspects that constitute this dimension remains a contentious issue among researchers. This ambiguity complicates the development of standardized models for vocabulary assessment, as highlighted by Schmitt and Meara (1997) and Schmitt et al. (2019). Consequently, designing effective vocabulary tests requires careful consideration of various factors, including test purpose, word selection, and the dimensions of vocabulary knowledge to be assessed.

Schmitt (2000) proposes a framework with four essential questions to guide test development: the test's purpose, the selection of words to be tested, the dimensions of vocabulary knowledge to be measured, and the methods of validation. While Schmitt et al. (2019) emphasize the importance of understanding the test audience and context, Schmitt's comprehensive approach provides a solid foundation for test conceptualization and validation. In this research, we adopt Schmitt's framework as a guiding model for developing and assessing vocabulary knowledge, ensuring a systematic and rigorous approach to test design.

1.8.1. Why Do you Want to Test?

Schmitt et al. (2019) and Read (2000) assert that visible statement of test purpose is a key feature in more recent test validation theory and test development. Because, test validation process requires a clearly expressed objective to check whether a test measures what it is supposed to measure; test purpose determines the way test scores will be interpreted and used to comprehend students' language proficiency.

In the field of vocabulary assessment, Schmitt (2000) summarizes a number of test purposes. One of them is to gauge students' achievement, for example, what words students have mastered after a course of language instruction. One more reason is to place learners at appropriate class level. TOEFL vocabulary size tests determine the extent to which test takers are proficient enough in a language to be enrolled in commercial courses. A test might also diagnose test takers' points of strength and weaknesses as regards lexical knowledge. Other reasons of testing lexis include fostering motivation to learn new words; eliciting learners' vocabulary learning development; and highlighting the importance of acquiring some words in a language.

All things considered, the purpose of this evaluation process is to assess learners' word knowledge after a definite language course, as such it is long term achievement test (attainment test). Achievement tests seek to elicit "how learners can use the vocabulary they have learned" Nation (2013, p. 522). This leads us to include an aspect of productive vocabulary use in writing. Now that the current test purpose have been delineated, the next section will be devoted to a description of word selection process.

1.8.2. What Words Do you Want to Test?

Selecting appropriate lexical items for vocabulary assessment is crucial and should align with the test's purpose. Nation (2013) emphasizes that word selection can be based on dictionary-derived methods or frequency lists. He suggests that the dictionary used should be sufficiently comprehensive to include all intended words. With the advent of corpus-based linguistics, vocabulary size tests often rely on established word lists, such as Coxhead's Academic Word List (AWL) and the British National Corpus, for word selection and grading purposes. For instance, Schmitt (2000) posits that most learners, regardless of their educational background, are expected to be familiar with the most

frequent English words. He further notes that for lower-level learners, frequency lists up to the 10,000-word level are suitable, while advanced learners may require sampling from a broader range of words.

The Common European Framework of Reference for Languages (CEFR, 2001) provides test developers with four strategies for selecting words to assess:

1. Choose key words and phrases relevant to learners' communicative needs and cultural contexts.
2. Select high-frequency words from extensive word counts or those pertinent to specific themes.
3. Utilize authentic spoken and written texts, assessing the vocabulary they contain.
4. Allow vocabulary development to emerge organically through learner engagement in communicative tasks.

In this study, principles 1, 2, and 3 guided the selection of lexical items. Words were chosen to align with thematic areas such as Ancient Civilization, Ethics in Business, Education in the World, and Feelings and Emotions, as outlined in the targeted textbook units. This approach ensures that learners engage with vocabulary relevant to real-life contexts. Additionally, authentic written texts were selected to provide prompts, offering learners opportunities to use words in context and produce original written work. Regarding achievement tests, Nation (2013) asserts that when assessing vocabulary knowledge for grading purposes, the sample of words should be based on those learners have been exposed to during the course, ensuring alignment with the instructional materials used.

Obviously, since the purpose of our research is to measure the achievement and quality of lexical knowledge, word choice will depend on the material covered in the classroom. Thus, using samples from word accounts is not possible as the principled objective is not to gain an estimate of vocabulary breadth of learners rather more stress will be put on learners semantic, morphological, syntactical, word use knowledge. To have a full understating of how these aspects have been favored, the next section will explain the whole procedure.

1.8.3. What Aspects of these Words Do you Want to Test?

The second step after word selection is to determine what aspects of word knowledge to test. The multi-dimensional character of lexis together with test purpose will dictate what features should be integrated in the test. Accordingly, assessing vocabulary is more an incremental process. Indeed, testing all aspects of a word is critical and relative even to proficient testing researchers of language. The test will integrate some of the aspects listed underneath. It is important to mention that these listed aspects are not exclusive to the present vocabulary assessment. These aspects are adapted from previously reviewed literature specifying dimensions of vocabulary knowledge for testing and teaching, especially Nation's framework of *What is Involved in Knowing a Word*.

-Lexical considerations: The form of the word; the written form will be tested in particular, and not the spoken one, since vocabulary will be tested via writing;

-Semantic considerations: Basically, any sort of vocabulary tests, whether directly or indirectly, test meaning regardless of their intention. Yet, knowledge of conceptual meanings and inferring meaning from context (the prompts) is directly addressed in this research;

-Grammatical considerations: Knowledge of the usual syntactic relations existing between words like word classes and their functions, collocations and associations the word has with other words; and

-Stylistic and register considerations: knowledge and understanding of metaphorical, collocational language use together with topical and stylistic situations.

Within this regard, Henrickson (1999) stresses three dimensions of vocabulary development that help researchers to decide about the likely lexical aspects to consider:

- An emphasis on receptive or productive vocabulary knowledge;

- The degree to which lexical knowledge aspects are controlled; and

- The degree of mastery (partial or precise) the test seeks to measure (by partial knowledge is meant to measure for example word meanings in one context not in all possible contexts).

This section is a preliminary first step towards applying Nation's (2001, 2013) framework. Given that, as Schmitt and Meara (1997) state, it is possible to use word knowledge as a research rationale; it is not possible to design a test all-inclusive of the word knowledge aspects. In a nutshell, when considering the above issues, aspects of word knowledge to approach are basically related to word meanings, word forms, and word use (semantic, morphological, syntactical and thus communicative aspects) in terms of depth aspects and productive knowledge in terms of dimensions; and how this knowledge is elicited is via writing in communicative contexts. The resulting test format would be named controlled productive depth of vocabulary knowledge.

As a necessity to incorporate language communicativeness, it is worthy to include context in addition to item recognition in our vocabulary test. We adhere to VDT for many reasons. We believe that a good vocabulary test is that test that measures learners' ability to correctly associate word forms with word meanings in addition to word use and this is not, however, the case for VST another vocabulary test that does not share the same characteristics. A tendency towards testing multiple aspects or levels of vocabulary knowledge, including measuring non meaning types of word knowledge (e.g. collocation, derivatives) is encouraged by various researchers (e.g., Laufer & Goldstein, 2004; Schmitt et al., 2019) to yield in informative findings about learners' depth of word knowledge. The overall aim of this research is to measure the "level of mastery" (Schmitt et al. 2019, p.3) of the three aspects of productive depth of vocabulary knowledge, namely, word meaning, word forms and word use.

As test purpose, word selection and what levels of vocabulary depth to be tested have been determined, now we look for more convenient procedures for how to test and approach this kind of knowledge as various measures have been established in the literature of vocabulary assessment.

1.8.4. How will you Elicit Students' Knowledge of these Words?

In this section, we explore the issues that underpin vocabulary test design and development (and the tasks we use henceforth). How to elicit learners' degree of mastery of target words forms a basic question in test development which is very easy to frame but difficult to answer. Because vocabulary knowledge is multidimensional,

assessing lexical competence in a single testing procedure becomes much more fuzzier, that is why it is suggested to use multiple tasks and contexts to provide a comprehensive image of learners' lexical knowledge and ability (Daller et al., 2007, p. xii).

Similarly, on the complex nature of vocabulary assessment, Webb (2005) and González-Fernández and Schmitt (2020) also assert that as research on vocabulary is rapidly increasing it demonstrates how complex vocabulary knowledge and acquisition are. Likewise, Schmitt et al. (2019, p.4) confirm that "it should not be surprising that testing this complex knowledge will require more sophisticated and precise vocabulary measurement, and more extensive literature surveys will be necessary to provide this". Years before, Read and Chappelle (2001) also condemned about the complexity of vocabulary measurement especially that there is no consensus about the overall test format for test development. Justification for this, commenting on vocabulary tests used for instructional or research purposes, they assert that there exist no standard procedure used neither for evaluating the existing assessment methods nor for creating future lexical measures.

Under these circumstances, this section is an attempt to find out a comprehensive measure of productive depth of vocabulary knowledge; even its conceptualization is not agreed upon and hence its assessment. Perhaps the most comprehensive framework of vocabulary assessment is Read's (2000) and Read and Chapelles' (2001). In thier framework, the authors posited that L2 vocabulary assessment principally underpins two different perspectives; the first viewpoint states that words can be tested as a semantic field out of context. The latter view indicates that vocabulary items, when measured, must always be in context. Pointing out to these two jointly contending viewpoints on vocabulary assessment, Read lists three vocabulary testing dimensions that are figured out in Table 1.4 (p. 59).

The three theoretical underpinnings of Read's framework (discrete vs. embedded, selective vs. comprehensive, and context –independent vs. context dependent) are further elaborated in 2001 by Read and Chapelle based on Messik's (1998) validation theory, who provides us with a framework of reference for systematizing L2 vocabulary assessment in our study. It is composed of eight vocabulary measures (Table 1.4) these

measures are described in terms of design, analysis and purpose. This framework incorporates all the possible uses of vocabulary tests.

Discrete A measure of vocabulary knowl- or use as an independent construct	⟷	***Embedded*** A measure of vocabulary that forms edge part of the assessment of some other, larger construct
Selective A measure in which specific vocabulary items are the focus of the assessment	⟷	***Comprehensive*** A measure that takes account of the the whole vocabulary content of the input material (reading/ listening tasks) or the test taker's response (writing/speaking tasks)
Context-independent A vocabulary measure in which the test taker can produce the expected response without referring to any context	⟷	***Context-dependent*** A vocabulary measure that assesses the test taker's ability to take account of contextual information in order to produce the expected response

Table 1.4: Three Dimensions of Vocabulary Assessment (Read, 2000, P. 9)

As the processes of word selection, item format and test length are three challenging and tricky issues that interfere when designing any vocabulary test, we find it worthy**,** comprehensible and workable to use Read and Chappelle's framework showing the nature of vocabulary tests as a validation model towards our study design. It guides us to the right selection of a vocabulary measure that will succinctly elicit the construct under study. This comprehensible proposal serves as a basis for developing new vocabulary measures. Described in three dimensions, this framework provides the design options (dimensions) for test development. It is also an instrument that aids the researcher to set an analysis for the current test purpose and how it can be systematically tied to test design and development; relationship between test design and purpose. It also claims that there exist a number of issues that should be regarded in the process of test design. These are to be considered in Chapter 4 in the study design.

From the abovementioned able (Table 1.4), it follows that Read (2000) suggests three continua of vocabulary assessment whereby word knowledge measurement has been viewed from three distinctive perspectives: The first design measure, known as

discrete vs. embedded*,* appears to approach vocabulary as a decontextualized set of items or simply as "discrete construct" (Read & chapelle, 2001, p.4). A clear exemplary case is illustrated in the assessment of individual content words in VSTs by means of definition, word recognition tests, MCQ tests …etc. Researchers adopting such approaches "treat vocabulary as a separate component of language knowledge, which can be investigated without reference to the functions of words in grammatical structures, text or discourse" (Read & Chapelle, 2001, p.2). The best example following this design option is the VST (e.g., Nation, 1983, 1990; Schmitt et al., 2001), an instrument used to measure some aspect of word knowledge of target words the aim of which is to provide an estimate of leaners' vocabulary size. It is clear that this type of tests emphasizes lexical knowledge of content words (form-meaning connection) and does not account for grammar and other language skills and even language proficiency is excluded because they are not contextually bound.

Contrary to the first dimension (*discrete*), the second dimension (*embedded*) is not strictly limited to the assessment of decontextualized content words but these are expanded in some sense. Therefore, the second measure of vocabulary assessment categorized under the first option headed ***discrete vs. embedded*** treats vocabulary as being an embedded component serving as a part of "the measurement of a larger construct" (Read & Chapelle, 2001, p. 4), be it writing proficiency. The ESL Composition Profile (Jacobs et al., 1981) is a clear case of this type. It is seen as a measure of writing proficiency that implements five rating scales one of which assesses range and appropriateness of vocabulary use. As such, vocabulary is rated distinctly as a separate component than joined to the other four scales to have a total picture of students' writing production (Read & Chapelle, 2001).

Read's second dimension of vocabulary assessment draws a borderline between ***selective*** and ***comprehensive*** measures. The former places more emphasis on individual content words extracted, for instance, from a reading passage whereby test takers' response to the test task is based on a comprehensive analysis. One example for the selective measure involves multiple-choice format that contains lexical items sampled from a written text and used as stimulus words companied by multiple-choice items.

Measures of lexical density is one case of comprehensive measures, which aim to compute the proportions of content words used by respondents in written or spoken test. This perspective approaches vocabulary as part of the learners overall proficiency in which lexical items are tested by means of performance tasks where production is fostered.

The third dimension distinguishes between ***context dependent*** and ***independent*** options. Here Read puts more stress on the role that linguistic context can play in L2 vocabulary assessment. In a context-independent measure target word are tested in isolation and learners are required to select responses from non-contextualized items; without reference to any linguistic context. In a context-dependent test, however, leaners' ability to properly use target words is assessed in context-bound settings, where learners have to demonstrate their lexical competence via texts that they would generate themselves.

Read and Chapelle (2001) sum up the three dimensions of L2 vocabulary assessment in eight vocabulary measures organized in a duality manner exemplified in different test formats (designs features). These are put forward in table 1.5 bellow:

Test	Features		
1) Vocabulary Levels Test	Discrete	Selective	Context independent
2)Lexical Frequency Profile	Discrete	Comprehensive	Context dependent
3) ESL Composition Profile	Embedded	Comprehensive	Context dependent
4) TOEFL vocabulary items	Embedded	Selective	Context dependent
5) Multiple-choice cloze	Embedded	Selective	Context dependent
6) C-test	Discrete	Selective	Context dependent
7) Vocabulary Knowledge Scale	Discrete	Selective	Context independent
8) Lexical Density Index	Embedded	Comprehensive	Context dependent

Table 1.5: Design Features of the Eight Exemplary Tests (Read & Chapelle, 2001, P.6)

Note. The tilde (-) symbol is used to indicate that the TOEFL vocabulary items are variably context dependent.

Elaborating on Read's (2000) three vocabulary assessment dimensions, Schmitt (2000) and Read and Chapelle (2001) argue that in the process of test design it is highly important to make specifications (decisions) about vocabulary design measure. Would it be tested as a decontextualized construct apart from listening comprehension and reading comprehension? How lexical items are to be selected? Furthermore, making decisions on the degree of linguistic context adhered to test input and targeted response are amongst the basic questions asked in the process of test design and development. Based on these questions, the coming section will provide a set arguments for the design measure opted in this study and the test format selected for testing depth of productive vocabulary knowledge.

1. 9. Framework of Vocabulary Assessment

This chapter was an attempt to establish a solid ground for test design and development, specifically in terms of content specifications and test format. The following are a set of justifications set forth for the current measurement procedure:

Justifications for opting for the third vocabulary measure:

- As far as the research aim is concerned, and perusing Read's and Chapelle framework (2001) of vocabulary assessment, selection of test format falls upon embedded, comprehensive and context-dependent continuum design option. The present PDVKT (Productive Depth of Vocabulary Knowledge Test) is an embedded instrument as it involves evaluating vocabulary via writing activity by means of which vocabulary is not treated as part of a general construct, but it is the construct per see. Vocabulary depth is to be elicited by means of writing short constructed responses (paragraphs) in relation to situation prompts and stimulus words. Unlike those measures developed in quest of lexical richness and sophistication, priority, when scoring leaners' writing performance, is given to vocabulary and its appropriate use but not writing proficiency.

- Comprehensiveness in test design entails that a test is context dependent and text-

bound. According to Read and Chapelle (2001) the text offers context for words combining it, and whether this text promotes comprehension or production, leaners are restricted to pay attention to context to create appropriate responses. In view of that, our test is an attempt to engage learners in production of appropriate responses with relevance to situations that are thematically-geared. Learners are required to use target words with their corresponding word-forms (word parts) in appropriate contexts suggested by situation prompts containing four topics: Ancient Civilizations, Education in the World, Ethics in Business and Feelings and Emotions covered in the textbook "*New Prospects"* devoted to teaching third year secondary school learners.

- As far as context-dependence is concerned, Nation (2013) suggests that words can be tested in isolation as they can be tested in context. This Context may refer to sentence context (decontextualized sentence context) or text context (passage context) and even exceeds to communicative context the point of interest in this particular research. He defines communicative context as being "an extended text where interpretation of the words is dependent on the wider context" (p.548). Testing vocabulary in realistic situations should expose learners to larger context that provide communications likely to be encountered in everyday interactions.

In Nation's (2001) terms, the current study explores two major issues: how well a lexical item is known and how well it is used. To this end, he claim that vocabulary should be integrated within the four language skills of listening, speaking, reading and writing to examine whether shortage in word knowledge is the direct cause of poor performances in the intended skills. Still, our aim is to explore respondents' quality of vocabulary knowledge and poor performance in lexis and not in the four language skills themselves, eventhough lexis is tested via writing.

Justifications for PDVK test

- Unlike grammar, vocabulary learning does not occur in a linear fashion in which students acquire more and more fixed rules or definite structures. Rather, vocabulary learning is a constructive process in which knowledge of new lexical items is not simply added but integrated into existing items (lexical map or repertoire threshold). During the learning process, words are rearranged in association with appropriate meanings that

is how meaning of the world is being constructed. Hence, this test is meant to assess how well leaners know how to create appropriate organized sets of words and put them meaningfully together within a written text.

- Research has revealed that vocabulary richness in writing and the proficiency level of students are two key factors totally linked to assessing the quality of writing (Nation, 2005). Through writing, test takers will show their productive use of language. For Nation (2013), writing activities elicit how far students' receptive knowledge (items in the input) is transferred into productive use (items in the written output). This conception inspired us to select words leaners were exposed to during a course of instruction. These words are assumed to be part of leaners' passive vocabulary and through the assessment we investigate whether they become part of learners' productive use ability via designing tasks for vocabulary performance in writing.

- Vocabulary items in a reading and listening assessment focus more on inferring meaning from context, it is mainly about meaning recognition (receptive knowledge of words), unlike assessment of words within the writing process where learners are required to assess the ability to use words in a purely original constructed response, be it a paragraph. That is, the productive knowledge intended is test takers's ability to use affixation to create new words and use them to produce novel meaningful sentences put together within communicative contexts. Zhong (2016, 2018) posited that short constructed answers such as paragraph witting would gauge learners' lexical performance, rather than writing a single sentence, because the latter process might yield in neutral sentences with the target word, such as, in 'expensive' that might be 'It is expensive' a sentence that does not reveal good command of vocabulary knowledge. Hence, testing PDVK via writing gives rich evidence of learners' lexical fluency.

It is worthy to mention that productive vocabulary is not tested via speaking as the course outcomes for the intended population focus more on writing ability other than speaking. The proof is that in the Baccalaureate (BAC) standardized assessment candidates are asked to write essays and not to produce spoken messages.

- All things considered, Schmitt et al. (2019) assert that,

Performance on a vocabulary test should be related to some kind of language use, such as

> listening, speaking, reading, or writing. Ideally, there should be some research data to support this, such as comparing test answers to some kind of measure of language performance (e.g. explaining the meanings of words when given the written word form, or answering a comprehension test of a listening passage in which the target words were embedded. (p.5)

In a nutshell, in L2 vocabulary assessment test designers, whether teachers or researchers, are required to specify test purpose and what words to test and how to test them. Besides that, they should specify the context of the assessment and the test format appropriate for testing purposes. Following this line of thought, the current PDVK test purpose is to assess learners' vocabulary achievement; in other words, students' ability to use target words in realistic contexts with their appropriate word forms (using word formation processes) and meanings. The test is assigned to first year university students enrolled in ENSB. It measures their vocabulary knowledge within the context of task-based paradigm.

This research derives from the existing literature taping different tests formats to achieve study aims of exploring the relative effects of task diversity in test results reliability. Paul et al. (1990) assert that "the choice of test formats depends on the type of information desired" (p.1). According to the researcher's interest in assessing vocabulary knowledge within the context of performance-based instruction as well as task-based approaches, the type of knowledge desired for testing made the choice of tasks and vocabulary use in situations. This is so being said, the aim of the current study is not to gain an estimate of total words that university entrants could know but to investigate, among other things, their productive vocabulary or lexical ability, mainly using words in context.

Conclusion

This chapter has discussed so far the contribution of lexical knowledge to the four language skills and to overall language ability. The reviewed literature proved that lexis is a very difficult field to investigate. Vocabulary is not simply that inventory of words neither it is an orthographic spaced set of letters; nor it is that knowledge of lexical functional structural semantic morphological properties of English language items

rather it is that complex entity of sub-knowledge encompassing form, meaning and use underlying the duality of receptive and productive knowledge. Theses knowledge types compile with Nation's aspects of knowing a word (DVK) that sound to be significant within vocabulary research realm. Word knowledge and depth of knowledge dimension in particular, represents a challenge for test developers when regarding the levels of competences as key points in test development. Aspects of word recognition go further conceptual meaning to embrace knowledge of word forms, formulaic language chunks, syntactic, collocational, stylistic and register constraints of lexical items. Other issues tied to which measure the researcher will adhere when developing a test for productive depth of vocabulary knowledge has also arguably proven to be challenging obstacles considered in the chapter.

This chapter argues for the purpose underlying measuring depth of vocabulary knowledge productively. PDVK is important to the students' writing and speaking overall proficiency. Unfortunately, however, the research reviewed indicated limited attention to vocabulary use in contexts with the aim of investigating the quality of that knowledge (it has not yet been researched to the best knowledge of the researcher) in large written discourse context to cater for the requirements of communicative endeavor. Measures used in the assessment (test batteries) of deep word knowledge are less extensively developed in vocabulary study, and even those well-established measures seem to test words in isolation regardless of context. The latter is also limited to the sentence level neglecting the larger communicative contextual aspects of vocabulary items. Especially that FLLs are to be faced with writing and speaking demands at university level. Hence, the purpose of the current work is to investigate EFL students' quality of lexical competence on three types of word knowledge, namely word meaning, word formation, and word use originated from Nation's famous framework.

In this chapter, we also considered mandatory issues in vocabulary test development and suggested that test development requires researchers and teachers to take into account testing purposes, word choice, what aspects of words to emphasize, word contextualization, choosing appropriate model to decide about test format based on

Read's famous framework of vocabulary assessment. Justifications for using word knowledge as a research rationale together with the steps applied for test design and development have been further discussed.

Now that the subject matter of the present research has been delimited, the next chapter will be devoted to performance-based assessment, a research paradigm set up for the validation and assessment of PDVK in larger communicative authentic contexts.

CHAPTER TWO

QUALITY CRITERIA FOR PERFORMANCE/COMPETENCE ASSESSMENT

Introduction

In this chapter, before turning to matters related to performance assessments, an overview of the defining principles of performance and performance assessment is provided considering its origins, conception, types and crucial characteristics. This chapter also sketches out the different methods and validation procedures used for assessing performance from initial design to developing a holistic scoring system to weigh examinees' product. It highlights the fundamental issues guiding performance assessment all-encompassing conventional quality criteria of validity and reliability accompanied by more new authentic forms of assessment, namely performance and competency assessments, shifting away from objective testing conditions.

In this chapter, an in-depth description of validity with particular emphasis on Messick's unitary approach to construct validity included together with its six aspects of validity evidence in addition to other modern quality criteria characterizing performance are discussed as measures for validation of the current performance assessment. These are elaborated in a comprehensive framework of quality criteria used for evaluating performance tests differently to standardized objective testing principles, emphasizing, among other things, authenticity, cognitive complexity, contextualization, meaningfulness, transparency, educational consequences, transfer and generalizability, cost, effectiveness, fitness for assessment purpose, and communicativeness. Finally, after the above criteria are reviewed in relation to the validation processes, a scoring guide, which is adapted from different sources, is suggested whereby assessors can rate test takers' performances. This chapter also explains the relationship between validity and reliability in G theory framework, where the latter is used to investigate reliability and validity (assessment precision) of performance/competence assessments by means of estimating the sources of measurement error. The closing section of this chapter highlights sources of variance in performnce and competence assessments; having a

significant impact on authentic assessment precision, these sources contribute to measurement error and score variability.

2.1. Origins of Performance Assessment

Performance assessment is a conception that gained interest during the last three decades. According to Lai (2011), performance-based assessments (PBAs) are antediluvian in scholastic/educational measurement endeavor. This is supported by Johnson et al. (2009) who see that assessing performance is not a holocene phenomenon. It can be traced back up to the Chinese culture, but what is current is to look for adequate methods to boost performance assessment together with its scoring systems. The reason why PBAs are prevailing-widespread currently is twofold: "to address the federal government's requirements for assessment systems [in America] that represent the full performance continuum" (Lai, 2011, p. 1) and to react to the limitations of standardized multiple choice testing.

What seems to come out of the MC testing practices is a disadvantage related to its maximizing of error through guessing and the fact that examinees are prescribed to select one answer from a set of distracters in the MC format, by no means, this "inhibits examinees from expressing creativity or demonstrating original and imaginative thinking" (Osterlind, 2002, p.164). As knowledge is not simply a matter of guessing or only simple statements; it is not limited to choice of alternative, it is rather shaped by means of mental abilities investement.

Performance appraisal, therefore, has been boosted in the American context and has become widely practiced all over the world ever since (e, g. Australia, Algeria, Malaysia, etc.) in many domains such as, music, sports, and education in general. In such assessments, test takers "demonstrate some type of performance or create some type of product (e.g., performance, performance-based, "authentic," constructed response, open-ended)" (Lai, 2011, p.1). Examples of this type that can be found in the literature involve:

-finding value judgments which invoves expressing or forming an opinion based on someone's' principles and beliefs but not on facts that can be proved.

-reasoning skills such as establishing relationships between situations and patterns such as establishing cause-effect relationships, or deriving some new information from existing information like explaining an issue or situation standpoints.

-application of certain previously acquired knowledge in new situations.

These were few examples illustrating performance, we presently limited our concern with it to breif examples for clarification purposes and in the sections coming next, we shall focus on performance and performance assessment conception in the educational literature, performnace types and performnce assessment characteristics.

2.2. Performance and Performance Assessment Defined

Evaluation and measurement specialists describe assessment as the process of gathering information about students' learning. Assessment is broader than testing and measurement because it includes all kinds of ways and procedures to sample, observe and obtain information about students' skills, knowledge and abilities or put simply performances. In this research, the overall focus is to assess the assessment, the measuring procedure, besides assessing students' vocabulary knowledge and ability. That is, students' performances, and their observed scores henceforth, are given a second priority in this particular study. Assessing students' vocabulary performance in a productive depth of vocabulary knowledge (PDVK) test is used as a procedure to obtain information about students' quality of vocabulary knowledge, or statistically observed scores, to assess the current assessment and not to grade learners or place them at particular groups. Hence, their performance is used to draw inferences and their observed scores will be interpreted from generalizability perspectives, namely reliability and validity of test scores.

As far as performance is concerned, Fitzpatrick and Morrison (1971) state that it is "a sequence of responses aimed at modifying the environment in specified ways" (p. 239). This "sequence of responses" involves all kinds of behaviors performed by an examinee. This refers to, what s/he can say, do or act out, or innovate (Messick, 1994). A "sequence of responses" points to the examinee's acts and behaviours; what they can write, say, or create. For "modifying the environment" a student can turn a piece of writing into a different form of literature, can transmit a written play into an actual text

performance with relevant theatre decoration, clothing, choice of actors, etc. Another example, a student's performance can be elicited in writing a chart of ethics in business after gathering information in law stream.

The simplest and succinct definition of performance might be provided by Gitomer (1993) who pens, it is "the execution of an action" (p. 244). Performance in this sense is restricted to any sort of behavior or action done by examinees in response to a given situation. In a language class, an examinee might act out a dialogue in response to an oral instruction, role play a dialogue in front of peers, tell a story to her/his mates, transfer information from a map into a manuscript giving directions to a foreigner in a new city, or derive new information from existing knowledge using inference, deduction or logical reasoning, etc.

A glance at the account that the aforementioned descriptions offer, when linking performance to our context of assessment, reveals that performance refers to what learners can do with the language in general and/or with the vocabulary learned in particular. In an attempt to "modify the environment", learners might be involved in the use familiar decontextualized words to express new thoughts incorporated within a given theme embedded within a system of ideas. These target words have been previously learned and are assumed to be part of students' repertoire. Learners can also be engaged in writing short constructed responses with relevant content especially pertinent to the task context situation, appropriate meaning associations, word meanings, sentence meaning, and above all discourse meaning. By "sequence of responses", in this research endeavor, is meant what students can write, produce, and create in relation to communicative tasks/situations using English language.

In more learner-centered approaches, performance-based tests involve oral production, written production, open-ended responses, integrated performance, group performance, and other interactive tasks rather than paper-and-pencil selective responses. The term might be best described by Dorfman et al. (1995), who insist that "an alternative approach to assessing students' achievements in school, refers to assessment methods that allow students to demonstrate their skills, knowledge, behaviour, and accomplishments across a wide variety of classroom domains on

multiple occasions". Performance assessments are those constructed responses created by examinees in response to a given prompt, task or situation and judged by qualified or proficient raters (Mc Bee & Barnes, 2009, p.180). As such, they involve learners in constructing their own meanings of the world unlike traditional assessment wherein learners have to select responses from ready made statements, they are different in nature and processes of answer construction.

It is obvious in performance assessment that tasks must be included as procedures to elicit students' performance. Performance assessment is task-based and examinees are required to demonstrate their knowledge and ability to apply that knowledge. A performance assessment "may comprise one or more tasks. A performance task is any activity that asks students to do something to demonstrate their knowledge, understanding, and proficiency. Performance tasks yield a tangible product and/or performance that serve as evidence of learning." (NASBE, 2020, p.35). In the language classroom, students are required to write paragraphs or essays, tell stories, act out dialogues, make oral presentations and conferences in front of peers, etc. These various performances might involve individual, pair or group work activities. For example, a performance task can be designed to determine examinees' competence to use learned/acquired words to solve a new problem, or to produce a written passage containing those words in a productive controlled vocabulary communicative task. Performance tasks, therefore, can be product, performance or process grounded. Assessing performance within the context of performance-based assessment lies in constructing a number of these task types that can be used to elicit students' performance and these are further described next.

2.3. Types of Performance Assessment

McTighe and Ferrara (1998) suggest three types of performance-based assessment: Products, performances, and process-oriented assessments. A product entails what learners can produce demonstrating concrete instances of ability to apply knowledge. Examples of this type can be found in various classroom assignments, whrere students are asked to produce reports, web pages, brochures, …etc. These might include homework being generally performed outside the classroom. In case

assignments are done inside the classroom under teachers' direct observation, these refer to performance-based assessment in which learners demonstrate their ability to apply knowledge and skills. Much of or all of the work might be done outside or inside school, but performance is suggested to be done inside the class where teachers and peers directly observe the results of their efforts and accomplishments. Performances, the second type of assessment, might comprise role plays, individual or pair oral presentations, demonstrations, and class discussions. In process-oriented assessments, students show their thinking and reasoning, and motivation in response to a given situation where students are required to reflect on their own learning and articulate their goals to improve it. They include think-aloud protocols, self or peer assessment grids, checklists or surveys, single or pair conferences.

This typology of assessments indicates how performance assessment methods are highly cognitively demanding and how they are authentic. They foster critical thinking skills and linguistic abilities as they bridge the gap between instruction and real world performance which are key factors towards assessing performance whereby learners have to process information and show a kind of processes used to respond to a given stimulus instruction.

Moreover, Fitzpatrick and Morrison (1971) distinguish between performance and product evaluation. A product form of performance assessment elicits students' production; it focuses on the outcomes of learning, or put simply, what learners can do as a result of instruction. Examples of this type involve producing essays and paragraphs. Performances, on the other hand, emphasize processes of learning. Performance evaluation often measure learners' progress via storytelling and conferencing for example.

McTighe's and Ferrara's (1998) classification of performance-assessments sounds exclusive, but that of Fitzpatrick's and Morrison (1971) seems the most applicable in educational research. Because, "[t]he categories of product and performance assessments offered by Fitzpatrick and Morrison (1971) encompass the various types of performance assessments used in educational and credentialing testing, research, and program evaluations" (Johnson et al., 2009, p.8-9). Accordingly, performance testing

targets both the products and processes of learning. “They encourage us to move beyond the ‘one right answer’ mentality and to challenge students to explore the possibilities inherent in open-ended, complex problems, and to draw their own inferences” (Herman et al., 1992, p.6). This new insight suggests to boost students’ performnaces when responding to questions that are often complled to different interpretations, where their product, performance, and cognitive skills and processes are assessed by expert judgement.

Overall, this section aimed to have a deeper understanding of the type of performance relevant to academic acheievement, and thus far performance assessment types have been described, the coming section is devoted to a detailed discussion of the various characteristics of performance testing tapped into the literature.

2.4. Characteristics of Performance-Based Assessments

Performance assessments have been viewed as multidimensional, situational and complex. Parkes (2000) complains about the complexity of the use of performance assessments especially in terms of the scores obtained. He favours authenticity in alternative assessments, and grouped the characteristics of performance assessments into three categories: those linked to the tasks themselves, those related to the performance elicited from students on an assessment, and those related to the scoring system.

In the language assessment literature, Bachman and Palmer (1996) provide five task characteristics in performance testing: “1) Characteristics of the setting (physical characteristics, participants, time allocated for task completion); 2) Characteristics of the test rubrics (instructions, structure, time allotted for scoring, scoring procedure); 3) Characteristics of the input (task format, language of input); 4) Characteristics of the expected response (task response format, language of expected response); and 5) Relationship between input and response (reactivity, scope of relationships, directness of relationships)” (pp. 47-57). These criteria are relevant to task-based performance that emphasize the processes of task construction, administration and evaluation.

What can be noticed from the two above sources is that performance assessment characteristics vary from one author to the other. Therefore, a variety of taxonomies of

performance assessment characteristics have been established by different authors. They either share much or something in common with Bachman's and Palmer performance testing characteristics mentioned above. Of the many of these elaborated taxonomies only Wiley's and Haertel (1996), Norris et al.'s (1998), and Norris' (2001) are examined in brief for reasons of space, and the remaning taxonomies are summerized and listed in Appendix A.

For Wiley and Haertel (1996), the environment, or conditions under which the task will be performed should be emphasized in performance testing, including the physical environment, timing, tools, equipment, physical resources, and kind of information to be made available. They also comment that any communications directed to the examinees performing the task should include delineation of the goal and evaluation criteria, and the tools that might be used to perform the task. As it seems, these performance testing characteristics are criteria for task specifications that can be applied in designing performance.

Norris et al. (1998) insist that performnce testing requires expertised subjective judgements from the part of the rater, a facet that must rely on appropriate rating scales, needs to vary in terms of number to maximize fairness, as it needs to be well trained in the use of scales to yield realiable results. These characteristics tap task scoring citeria used for rating students' performance.

Norris (2001) fosuses on task and content selection of the performance test. The author describes the tasks to be constructed for university level English for Academic Purposes students. These tasks should be characterized by:

- General interest to university-level L2 English users;

- Representativeness of various content areas;

- Not being highly discipline-specific;

- engaging the examinee in a variety of complex, skills-integrated L2 activities; and

- Maintain real world dependability to the greatest extent possible. As can be seen, the researcher emphasizes characteristics of the scoring process.

In brief, the above criteria set for planning, developing, administering, scoring and evaluating performance tests underscore, among other things, high order thinking/cognitive skills, task completion, authenticity, complexity, meaningfulness, communicativeness and contextualization. In the process of test design and development sense, the above task characteristics, even called so, resemble task specifications, task content and criteria used for scoring students' performances. Overall, all the features of performance testing share fundamental things in common; they require examinees to respond to a given task or situation, seeking to elicit what examinees can do, produce or create demonstrating their knowledge and ability. In this sense, performance shifted away from that simple recall of knowledge to more constructed response prompting cognitive abilities and reasoning skills.

2.5. From Objective to more Performance/Competency Assessments

Since 2006, the Algerian education system, like many others globally, has adopted a performance and competence-based approach to education. Prior to this shift, assessment practices were dominated by objective testing—standardized formats such as MCQs, true/false, short-answer, and cloze tests. These were popular due to their practicality, ease of scoring, broad content coverage, and their perceived objectivity, validity, and reliability. However, these methods have been criticized for their inability to measure higher-order thinking skills like problem-solving, research, or project work (Ruiz-Primo & Shavelson, 1996). They also fall short in assessing key cognitive and metacognitive skills such as analysis, reasoning, and self-evaluation.

While Osterlind (2002) argues that well-constructed MCQs can assess complex thinking, achievement tests largely remained confined to factual recall and discrete skills. As Linn et al. (1991) highlight, such tests encourage superficial learning, limit curriculum scope, and promote teaching to the test. Allem (2000) further critiques standardized tests for overlooking higher-order cognition, fostering instructional narrowing, and failing to ensure equitable skill acquisition.

In response to these limitations, a paradigm shift emerged, favoring assessments based on performance and competence rather than rote knowledge. Researchers such as Baxter, Shavelson, Goldman, and Pine (1992) emphasize three driving factors behind

this shift: dissatisfaction with MC formats, advances in cognitive science, and curriculum reform. These changes reoriented assessments toward conceptual understanding, problem-solving strategies, and the application of knowledge in realistic contexts.

Performance assessment encourages tasks that demand higher-order thinking tasks that engage students in meaningful cognitive processes directly linked to classroom activities (Ruiz-Primo & Shavelson, 1996). Students are thus evaluated not just on what they know, but on what they can do with what they know. These assessments prioritize open-ended tasks and authentic applications (Jonsson & Svingby, 2007), aligning evaluation with real-life skills and practical competence.

From the late 1980s onward, calls for authentic assessment grew stronger. Wiggins (1989, 1993) and Keller et al. (2010) stress the value of assessing skills through tasks that reflect real-world challenges. McBee and Barens (1998) note that performance-based assessment provides high content validity and better reflects students' integrated abilities to solve problems.

In this context, achievement is understood as successful task performance. Allem (2000) broadens the notion of competence to include cognitive, comprehension, and complementary skills beyond what traditional methods can capture. Though space limits a detailed discussion, competence is viewed here as a practical construct, akin to Chomsky's (1965) linguistic distinction between competence (knowledge) and performance (actual use). Delory (2002) and Scallon (2002) treat competence and performance as interchangeable, arguing that both are validated through similar assessment criteria and conditions.

Thus, in competency-based assessment contexts, what a learner *does* reflects what they *know*, making performance the observable evidence of competence. For example, lexical competence is revealed through language use in tasks.

However, this shift has not come without challenges. Performance and competence assessments, due to their reliance on open-ended tasks, raise issues of subjectivity and reliability (Shavelson et al., 1993; Brennan, 2000; Parkes, 2001). Scoring variability, rater bias, and test-retest inconsistency complicate validation and

design. Moreover, increasing the number of open-ended tasks boosts validity but also raises concerns of cost, efficiency, and practicality.

Ultimately, while performance-based assessment offers richer, more authentic insights into student learning, its implementation demands careful attention to quality control, particularly in terms of objectivity, reliability, and design feasibility.

Overall, adhering to alternative assessment pushed researchers to find out new methods to envision and embody performance and competency assessment program evaluations. This practice has been reflected in the current attempts to tap into low reliability and validity of assessment systems being subject to measurement errors. In practice, efforts have been made to find out solutions to achieve high degree of quality criteria that can decrease cost and efforts, and provide new criteria for efficient evaluations such as, authenticity, cognitive complexity, credibility, costs and reform, etc.

2.6. Quality Criteria for Evaluating Performance/Competence

In an attempt to fit the new forms of alternative assessment that flourished in the previous century, instruction and learning embraced new measurement tools. There was a call for developing other methods that can cater for acquisition of competence. Because evaluating competence is a complex process, it needs more than one procedure and more than validity and reliability as quality criteria. Competency Assessment Programs (CAPs) have integrated new and various methods in evaluation (Baartman et al., 2006; Baartman et al., 2007) including open-ended tasks, performances tasks requiring short or long constructed responses, portfolios, etc. instead of single assessments. The classical psychometric quality criteria implemented in traditional assessment are outdated, still remain significant to be implicated again in CAPs (and performance assessment programs henceforth) but should be operationalized differently.

2.7. Classical Quality Criteria: Reliability and Validity

At the outset of the 20th century, psychometric theorists and practitioners investigated issues of test validation namely reliability and validity. These two issues

were regarded key criteria for assessing the quality of standardized MC tests. Reliability had received great attention in the classical psychometric framework and, more precisely in achievement tests (Linn et al., 1991) especially with the application of more standardized objective testing. Besides reliability, validity had also gained great consideration afterwards as a second feature of assessment quality criteria.

Some time after the rise of alternative assessment, new forms of assessment based on performance and competence as well as authenticity continued enormously to investigate issues of reliability and validity, and treated them as being inevitable criteria in any sort of evaluation. However, how important these criteria are was not the question of alternative assessment, rather how can they be implicated in the new assessment paradigm (performance/competence assessment programs). Another central question was seeking more information to check whether CAPs are of similar quality as those of classical forms (Baartman et al., 2006; Baartman et al., 2007).

Traditionally, reliability was conceived of as the consistency of measurement scores across occasions (repeated occasions), consistency of items or tasks (inter-task consistency), and inter-rater consistency. Estimating consistency depended on statistical indexes suitable for the nature of the facets the decision maker is willing to investigate (replication, agreement among raters, internal consistency). Whereas, traditional conception of reliability cannot be adopted in CAPs with different assessment regimes. Classical procedures conducted in objective testing to estimate reliability are not suitable to evaluate competences (Scallon, 2004; De ketele & Gerard, 2005), especially that the current assessment depends on criterion-referenced testing and complex open-ended tasks (Cronbach et al., 1997). In effect, there was a strong emerging need to look for alternative measurements, which undergo sufficient practical methods designed to assist assessment stakeholders in improving suitable judgments (Gipps, 1994; De Ketele & Gerard, 2005). So far, we have briefly considered reliability from the perspective of Classical Test Theory (CTT) and how far it is critical to performance testing. For more information on reliability within Classical Test theory and generalizability theory paradigms, see Chapter 3).

Another criterion which sounds rather decisive of assessment quality which has stemmed from conventional assessment is validity. Validity has been conceived of as the test ability to measure what is supposed to measure, i.e. whether the test truly matches the targeted objective. To ensure validity of assessment tools, three types of validity were highlighted in traditional assessment: Content validity, criterion validity, and construct validity. These types are usually evidenced using correlation coefficients. Recently, however, after the rise of new evaluation methods related to performance and competence assessments, the notion of validity shifted away from the usual meaning (Scallon, 2004). Please note that a full description of validity will be provided next in in this chapter in section 2.8.

Nevertheless, for convenience sake this does not mean that the notions of reliability and validity are wrong and misleading or they should be eradicated from performance assessment programs and CAPs, rather they should be investigated in different ways, and be reviewed in away suitable for further accumulation with alternative assessment (with multiple facets and specific characteristics). According to National Research Council (2002), the classical quality criteria of reliability and validity are crucial and should highly serve as basis for performance assessments. This insightful trend is much shared by authors like Kane (1992, 2004), Baartman, Bastiaens, Kirschner and Van der Vleuten (2006) who point out that fundamentals of CTT are likely to be implemented in quality assessment of competences. Similarly, Scallon (2004) adheres to the role of validity and reliability in tests evaluating competences, because both classical terms of validity and reliability are not wrong in basis but they suffer from certain problems.

CTT reliability, as will be revealed in chapter three, has proved to be weak, as its investigation was associated with the true score variance, carried out in a single facet analysis that treats one type of reliability index, including either test-retest reliability, internal consisiteny, or inter-rater consistency. It examines only a single error of variance at a time, neglecting other effects that might be affecting a measuring procedure. Adversely, reliability in performance assessment is examined from multiple perspectives where variance components (facets of measurement such as tasks, raters,

contexts...) are treated via G theory to estimate assessment precision that depend on the level of reliability index. Validity, on the other hand was conventionally viewed as validity of instruments, and alternatively in performance assessment is seen as the validity of inferences made on test scores.

Since reliability will be further discussed in chapter three, a brief account of it will be provided in section 2.10 discussing reliability within the context of performance-based assessments. Validity, however, will be worthly devoted a detailed discussion being a revisited conception witnessed past views and recents conceptions.

2.8. Validity Defined

Validity has always been regarded as a core feature in judging the quality of tests; a characteristic once evidenced in a test it is qualified as good. It is at the very heart of language testing and assessment research. The traditional conception of validity underpins the major question that was first posed by Lado (1961) in his book *Language testing*: "Does the test measure what it claims to measure?" (p. 321). The extent to which a test accurately measures the intended construct or test taker's abilities is a "one question approach" (Im et al., 2019, p. 3). It could be described as a trinitarian approach that, in the literature of psychometrics, utilized three forms of validity according to Shepard (1993). Content validity, construct validity, and criterion-referenced validity were the basic perspectives used to judge the trustworthiness of measurement procedures and are intended to accomplish different purposes (Messick, 1989; Shepard, 1993).

Content validity is used to address the extent to which test items adequately measure or cover the intended domain or construct. To illustrate, is the test content on a vocabulary knowledge test representative to the expected domain and defined language ability? Construct validity is used to determine the degree to which test score interpretations align with the construct in play; does the test of productive depth of vocabulary knowledge measure the construct that it is designed to measure; whether it measures learners' knowledge and productive use of language. Criterion-referenced validity maintains that when a test has criterion validity, this means that the scores obtained do accurately measure the concrete outcome that it is intended to measure. By

concrete outcome is meant, for example, learners' ability to comprehend target words, use appropriate word-formation processes to from new words, and use them in a variety of communicative contexts.

To cater for the changing needs of alternative assessments, the concept of validity was devoted many definitions in many psychological and educational measurement published articles. Validity is associated with measurement tools' ability to allow for making decisions related to certain assessment objective or objectives, this means that it is not classified among test characteristics. Validity, henceforth, has an inferential quality in nature (Messick, 1994). Validity from its broadest sense is a method of collecting information (arguments) to support the degree to which inferences or certain interpretations can be accurately made on students' performances (products) for evaluation purposes (linn, 1994; Moskal & Leydens, 2002). Similarly, decades ago, Cronbach (1971) described validation as the process used to collect evidence to approve the kinds of inferences that test developers and test users draw from test scores.

In a similar vein, validity as an important issue in test construction and evaluation is related to the test content, one feature of good tests addressed in traditional testing, and consequences of test use is another feature appeared in more performance-based assessments. Linn (1994) defines validity as "the adequacy and appropriateness of uses and interpretations of results" (p. 565). This quote states that the process of validation involves an evaluation of the impact of test use or what is known as washback in the literature of assessment. Crocker and Algina (1998), on the other hand, believe that validity refers to those inferences drawn from test scores about students' achievement in non-test setting. Likewise, in 1999, American Educational Research Association (AERA), American Psychological Association (APA), together with National Council on Measurement in Education (NCME) declared that validity is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). This view to validity has already been elaborated by Messick in 1989 in his article on *validity* and in his "*The interplay of evidence and consequences in the validation of performance assessments"* (1994) in particular.

In his famous quote that revolutionized the field of educational measurement, Messick (1989) penned, “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (p. 13). This quote suggests that validity alludes to the scientific arguments a researcher uses in order to approve whether types of inferences drawn from test scores are adequate and appropriate for decision-making whether norm-referenced or criterion-referenced. Scientific rigor is all-encompassing a complementary set of empirical evidence that is based on observation or experience and theoretical evidence based on theories and hypothesis. In his initial writings, Messick maintained that validity refers to “the relation between the evidence and the inferences drawn that should determine the validation focus” (p. 16). Concisely, validity is all about the confidence a test establishes; it is about whether a test assesses what it is designed to measure (Cumming, 1996).

Hence, test developers and users, according to Messick (1989), AERA, APA, NCME (2014), and Iliescu and Greiff (2021), should bear in mind that validity does not refer to instruments but to inferences. This is quite illustrative of the current conception of validity, as a synthesized entity of information collected from empirical evidence and theoretical reasoning, more precisely, it is that “evidence and theory [which] support the interpretations of test scores for proposed uses of a test”(AERA, APA, NCME, 2014, p.11).

All in all, validity, a core issue in assessing the psychometric quality of assessment was traditionally viewed as one of the assessment quality criteria of test scores, and of the quality of a test itself. In educational measurement, it is a method used for judging test usefulness and explaining test scores. In this sense, it is limited to the type of evidence it can provide for the decisions made based on test results. Throughout time, the concept of validity has brought a change in meaning; from being extremely one characteristic of good tests to an inference or evidence characteristic of observed scores. The term validity has been developed, after being more classical in measuring test

defined characteristics to being more effective in setting more suitable explanations and use of test scores.

From the above definitions, it can be assumed that the notion of validity is a complex construct indeed but interesting enough to be worth the attempt to grasp its various types, and highly relevant since it is inherent to vocabulary and performnace assessment; it establishes arguments for the adequacy and appropriateness of inferences and interpretations made on test scores. Hence, the section that comes next will be devoted to describing the types of validity back to the 1930s.

2.8.1. Types of Validity

Angoff's (1988) historically reviewed the practices of psychometric literature since 1930s whereby he provided a detailed account of divergent types or conceptions of test validity dominating the academic and educational scene at that time. He described validity as an evolving concept in 16 types. These are summarized next:

- Concurrent validity: “correlation of test against criterion; but here the predictor scores and the criterion scores were observed at the same point in time. Concurrent validity data were taken as evidence that a newly proposed test, or a brief version of an existing test, was measuring a given trait”, (p. 21)

- Construct validity: “a mutual verification of the measuring instrument and the theory of the construct it is meant to measure” (p. 29); “Cronbach and Meehl maintained that we examine the psychological trait, or construct, presumed to be measured by the test and we cause a continuing, research interplay to take place between the scores earned on the test and the theory underlying the construct”, (p. 26)

- Content validity: “a review of the test by subject-matter experts and a verification that its content represents a satisfactory sampling of the domain”, (p. 22)

- Convergent validity: “correlations among different methods of measuring the same construct”, (p. 26)

- Criterion-related validity: “combined concurrent and predictive validity and referred to the two as a class called criterion-related validity”, (p. 25)

- Factorial validity: “the loading of a particular factor on the test in question”, (p. 25)

- Discriminant validity: “correlations among different constructs that are measured by the same methods”, (pp. 26- 27)

-Ecological validity: “variation in validity coefficients across different (…) situations”, (p. 28)

- Face validity: “the appearance of validity (…) the tests should be constructed to face valid, that it is important, for example that the language and contexts of test items be expressed in ways that would look valid and be acceptable to the test taker and to the public generally”, (pp. 23-24)

- Intrinsic validity: “any given measure of a construct should show strong relationships with other measures of the same construct, but weak relationships with measures of other constructs”, (p. 26)

- Operational validity: “(…) defining the trait in question in purely operational terms as simply that which the test measures”, (p. 20)

- Predictive validity: “the correlation between observed scores on the test with observed scores on the criterion”, (p. 20)

-Population validity: “variation in validity coefficients across different populations”, (p. 28)

- Task validity: “variation in validity coefficients across different (…) tasks”, (p. 28)

- Temporal validity: “variation in validity coefficients across different (…) points in time”, (p. 28)

- Validity generalization: “studies of (…) effects [populations, tasks, situations] which have been referred to recently as yielding evidence of validity generalization (…) provide extremely important data and theory that have value not only in testing and modifying the construct, but in the strictly utilitarian sense of extending the contexts in which a predictor may be useful”. (p. 28)

The above list of validities is no exclusive, even daunting, as some other researchers twisted other sounding types of validities especially related to face validity. Lather (1986), for example, suggested catalytic validity as a measure of validity in

critical inquiry. She asserts that rigor and credibility of data are requisite for critical social research. For Lather, it represents the degree to which the research process reorients, focuses, and energizes participants toward knowing reality in order to transform it. Efforts to produce social knowledge that will advance the struggle for a more equitable world must pursue rigor as well as relevance.

Catalytic validity, according to Lather, alludes to the fact that research exhibits reality and empowers the research respondents to acquire a sense of self-understanding, activism and self-determination. To further ensure scientific rigor and credibility, she adds three other conditions:

- triangulation of methods, theories and data sources;

- reflexive subjectivity: the documentation or evidence of how assumptions have been impacted by the logic of the data;

- face validity: Face validity is established by recycling categories, emerging analysis, and conclusions through at least a subsample of respondents.

The aforementioned validities proved the psychometric conventional language test practices and notions that are, afterwards, replaced in modern times by other validation[4] criteria, often known in the literature of performance assessment as quality criteria of test design and development following Messick's Progressive Matrix of construct validity. The theme that this matrix generated is further discussed along the following lines.

2.8.2. The Centrality of Construct Validation/Validity

Informed by the scholarly work on construct validity, the research seeks firstly to reach a working definition of what construct validity stands for. Secondly, it attempts to arrive at greater understanding of the type of knowledge relevant to language and educational performance assessment. A question that might be asked here, is how different is this section from the above? In search of developing, assessing, and validating language tests and the investigation of the effect they may have on education and society, modern language testing theorists, test developers and test validation analysts (e.g., Bachman & Palmer, 1996) discussed and revised the concept of language

test validation resulted in developing innovative techniques for both assessing and validating language tests. Therefore, the above mentioned conceptualizations (the 16 types) of validation has been replaced by a central concept called construct validity by Messick in 1989. According to Cumming (1996), construct validity "has been widely agreed upon as the single, fundamental principle that subsumes various other aspects of validation (i.e. those listed above), relegating their status to research strategies or categories of empirical evidence by which construct validity might be assessed or asserted" (p. 5). As such, construct validity was widely adopted by language testing developers as a unified framework, and this is the case for our study concern.

Construct in research refers to the foundational knowledge, skills and abilities and other attributes that represent the main focus of the test. Construct validity is "an integration of any evidence that bears on the interpretation or meaning of test scores" (Messick, 1989, p. 17). Following Messick's progressive matrix of construct validity, Bachma (2003) defines construct validity as "the extent to which the results of an assessment can be interpreted as an indicator of the ability we intend to measure, with respect to a specific domain of generalization" (p. 7). From Messick's and Bachman's conceptions, it can be stated that construct validity incorporates any form of validity evidence including classical indices of content or criterion validity discussed before, because they contribute to the meaning of test scores interpretations. When construct validity is compared to traditional conceptualization of validity, it could be asserted that validity was considered a quality of the test per see; however after the conception of construct validity has been firstly introduced by Messick it has been taken to mean a quality under which the test is used. In other words, in this conception validity can be viewed as a steady property of the test itself, a property which is not threatened by other factors dictated by the context under which the test is used.

The construct validity theory, led by Messick (1989), asserts that the six aspects of validity or what he calls validity evidence are required to achieve an all-inclusive judgment concerning the validity of the use and the interpretation of results of a test, but not the judgment related to test content. Neverthless, content valididty can be used as evidence of construct validation being one aspect of validity used to prove that test

results are valid and appropriate to make decisions (e.g., grading learners, success or failure). Validity in such terms is applicable to the evaluation of the assessment quality rather than determination of the extent to which a test targets what is supposed to test. To assess the quality of the assessment or its validity, Linn (1994) considers issues of generalizability, fairness, and comparability besides Messick's aspects of construct validity, because they are inherent in validity.

Content, criterion, and construct validities were the core aspects of traditional conception of validity. In modern assessments, however, Samuel Messick (1995) in his article titled *Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning,* introduces his more comprehensive theory of construct validity that considers both score meaning and social values. Messick (1988) questioned the validity of validity and inquired whether it is satisfactorily representative of relevant content, or is entirely trustworthy in educational assessment practices. He divides construct validity into six componential aspects, namely, content, substantive, structural, generalizability, external, and consequential aspects that will be explored next.

2.8.3. Aspects of Construct Validity

2.8.3. 1. Content Aspect of Construct Validity

This aspect entails "evidence of content relevance and representativeness as well as of technical quality (e.g., appropriate reading level, unambiguous phrasing, and correct keying)" (Messick, 1996, p.9). It alludes to the degree to which content is specified and how much it delineates to the domain of the construct. Delineation, which refers to the extent to which content aligns with the construct under study, and structure of the construct domain can be addressed via task analysis, curriculum analysis, and especially domain theory. Simply, it entails how far the measuring procedure constitute all aspects of the construct. A test is judged of low content validity if it lacks some aspects or involves irrelevant factors. In our case testing aspects of productive depth of vocabulary knowledge via writing includes word meanings, word formation and word use; if word use is missing, the type of vocabulary knowledge tested is not productive and hence the test has low content validity. Irrelevant aspects might include

pronunciation because it is a paper and pencil test and examinees do not exceed to responding to an oral instruction where pronunciation and aspects of prosody can be measured. Validity is reduced because the meaningfulness and accuracy of test scores are affected by either irrelevance or missing aspects and thus for decisions to be made upon test findings.

All things considered, the content aspect of construct validity involves test specifications regarding test alignement with test objective; the degree to which test items are relevant. The results thus obtained are suggested to be useful to draw inferences and build interpretations upon them.

To achieve a high level of, or at least acceptable level of, content validity, we will address how well the current measurement is relevant to curriculum and course objectives by means of establishing a table of specifications (see chapter four for a full description). We also checked, in chapter one, whether the test content aligns with previously established research theories and concepts. A packed description of the concept of depth of vocabulary knowledge, and finally we came up with the construct of vocabulary being a set of componential aspects based on Nation's framework of what is involved in Knowing a word (see Chapter 1). The construct of vocabulary, therefore, was believed to satisfactorily incorporate word meaning, word forms, and word use. These factors were suggested to be a useful representation of the abstract construct we are trying to measure. The research aims to check whether the present measurement truly captures, or simply measures, all the dimensions of the construct intended to be measured. If so, the tset will have high content validity, if not the validity is thus low.

2.8.3.2. The Substantive Aspect of Construct Validity

In the literature of psychometrics, the substantive aspect of construct validity pertains to the validity evidence based on response process. It refers to the "theoretical rationales for the observed performance regularities and item correlations, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (Messick, 1996, p.9). According to Coulacoglou and Saklofske (2018), the substantive aspect indicates the extent to which theories and process modelling play an

important role in determining the domain processes expressed through test items. The domain processes is addressed by collecting evidence based on response processes. The standards (AERA, 1999) assert that "evidence concerning the fit between the construct and the detailed nature of performance or response actually emerged in by examinees" (p.12). Validity evidence collection based on response processes include identifying the cognitive strategies implied by test takers or excluding particular constructs-irrelevant strategies, like guessing or test wiseness (coulacoglou & Saklofske, 2018). To this end, a number of methods can be used to address the issue of domain processes including think-aloud protocols, interviews, and focus group such as group discussions …etc. In this regard, the construct boundaries or domain coverage does not only imply content representativeness of the construct domain rather it subsumes the processes representing the construct and how well these are mirrored in the measuring instrument.

Substantive validity emphasizes investment of students' responses to the test tasks to delineate the construct domain. This aspect will be addressed in the process of test development and validation during the piloting phase (domain processes) using a questionnaire and benchmarks.

2.8.3.3. The Structural Aspect of Construct Validity

According to Messick (1996), structural validity which is also known as internal validity "appraises the fidelity of the score scales to the structure of the construct domain at issue with respect to both number (i.e., appropriate dimensionality) and makeup (e.g., conjunctive vs. disjunctive, trait vs. class)" (p.9). This aspect indicates that the theory underpinning the construct domain should address issues related to the design and development construct–based scoring criteria along with selecting and constructing adequate test items or tasks (Coulacoglou & Saklofske, 2018). Accordingly, the internal structure of the assessment corresponds to the interrelations among the scored aspects of task i.e., scoring criteria, and task performance that, in turn, should correlate with the construct domain (Messick, 1989; Coulacoglou & Saklofske, 2018). Generally speaking, evidence based on internal structure, according to the *Standards* (AERA, APA, & NCME), concerns "the degree to which the relationships among test items and test components conform to the construct on which the proposed

test score interpretations are based" (p. 13). Thus, internal validity pertains to the assessment conditions; how well a test is designed, structured and conducted emphasizing the alignement of the scoring procedure with the content covered.

Andrade **(**2018) asserts that internal validity is concerned with an investigation of the way in which a study, be a test, is designed, conducted and analyzed leads to answers worthy of trust to the questions posed in the study. To illustrate, if a researcher adheres to unintentional unblinding of scorers, this would hinder the trustworthiness of the conclusions and inferences. Andrade further comments that other factors threatening the fidelity of the results in internal validity are related to improper randomization, which may generate missing data. Besides, internal validity aspect endeavors the presence of systematic error due to bias that may, for example, result from selection bias, and/or from performance bias (Juni et al., 2001, as cited in Andrade, 2018). Overall, structural validity examines if the study design answers the research questions with no bias. It is all about making judgments about the trustworthiness of the assessment procedures as concerns its structure and this can be achieved by training raters in the use of scoring guides.

There are three basic aspects of internal structure: dimensionality, measurement invariance and reliability (AERA, APA, & NCME, 2014). Structural validity "is the degree to which scores of a scale are an adequate indication of the dimensionality of the construct, attribute or factor being measured" (Brown, Bonsaksen, & Hui, 2019). Collecting validity evidence based on the internal structure source of validity can involve factor analytical studies, differential item functioning studies, and item analysis (AERA, APA, & NCME, 2014). Internal validity is based on judgments not statistical evidence. In this particular research, internal stduture will be addressed by training scorers in the use of established scoring rubrics, and it will arguably proved via a checklist of content validity of assessment tasks.

Several statistical procedures can be used to assess the internal structure of an assessment. The best method that fits into the evaluation of structural validity aspect in general and internal structure in particular is the Rash Measurement Model (RMM), a type of Item Response Theory (Bond & Fox, 2015). This mathematical/statistical mode

assumes that not obligatorily each item of a scale has the same value or duplicates the same difficulty level. The RMM identifies both students' ability and item difficulty generating a hierarchical scale of items; from easy to difficult. (Brown, Bonsaksen, & Hui, 2019). In this research internal validity is addressed by applying G theory to statisctically determine whether the measuring instrument can rank order items such as tasks and themes on a scale in terms of their level of difficulty simultaneously demonstrating participants' lexical knowledge and ability.

All in all, the test structure implicitly indicates the test componential parts are related to each other underpinning the entire theoretical construct.

2.8.3.4. The Generalizability Aspect of Construct Validity

This aspect which is also known as validity generalization "examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including generalizability of test-criterion relationships across settings and time periods" (Messick, 1996, p.9). According to the generalizability aspect, research results might be absolutely valid in one measurement setting but not in another. In this context, Messick (1989) looks at G theory from two different perspectives: G theory gives a natural framework for exploring the extent to which performance assessment findings can be generalized (Linn et al., 1991). One evidence for the generalizability of test scores lies in its reliability index achieved. That is reliability is prerequisite to validity; a reliable test is valid and hence the test results are generalizable. Messick (1996) considers generalizability an aspect of construct validity that could be interpreted in terms of reliability or transfer. Generalizability as reliability alludes to the consistency of performance across tasks, occasions, and raters of a given assessment which might very restricted in scope, while generalizability as transfer means "the range of tasks that performance on the assessed tasks is predictive of" (Messick, 1996, p. 250). Hence, the generalizability concept can be perceived in terms of either consistency of scores or transfer of test tasks to a larger domain, Bachman and Palmer (1996) call this domain "target language use."

In our study the generalizability of scores will be exmined in relation to the reliability index obtained from G studies and analyses.

2.8.3.5. The External Aspect of Construct Validity

External validity alludes to the validity evidence based on associations with other variables, more succinctly, “the extent to which the assessment scores' relationships with other measures and non-assessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed” (Mesick, 1996, p.12). For Andrade **(**2018), external validity investigates if the results of a study can be generalized to other settings. In this case, random sampling yields good external validity because, random samples are representative of the population of interest (from which a sample is selected) to which the findings can be generalized but not to other populations. External validity allows for a better evaluation of assessment; indicating how far the results of a study are generalizable to other contexts which means that the results are fidelitous.

Gren (2018) proposes that external validity evidence subsumes ecological, convergent, discriminant and predictive types of validity (see section 2.8.1 for a full description). Gren suggests a checklist of items related to external validity all-encompassing ecological, convergent, discriminant and predictive validities. The checklist includes four items set up under one emajor question: “does the test have ecological, convergent, discriminant, and predictive qualities?” More specifically, a) “ecological-is the real world behavior in accordance with how a subject answers the test? b) convergent- is the test similar to (converges on) other operationalization that it theoretically should be similar to? c) discriminant- is the test dissimilar to (diverges from) other operationalization that it theoretically should not be similar to? d) predictive- can the test predict something it should theoretically be able to predict?” (p. 3)

Besides, convergent, discriminant, ecological and predictive validity mentioned above by Gren (2018), Coulacoglou and Saklofske (2018) suggest that external validity also involves other traditional types of validity like criterion-related and concurrent validity that can be considered su-aspects of external validity (see section 2.8.1 on types

of validity). External validity is further described as validity "based on the relationships between test scores and other variables extends beyond single-test criterion relationships. This type of validity incorporates the analysis of the relationships of test scores with constructs that are expected to be related to positively or negatively or to be unrelated" (Coulacoglou & Saklofske, 2018, p.55).

External validity, including its subtypes of validities, is integral in conducting and analyzing studies. It is also known as the generalizability aspect because it examines the level to which test results are generalizable to other situations, tasks, forms, or to other facets of measurement in general. External validity in this sense emphasizes how applicable the test results and score interpretations are in the real world. It shows if the current study results are meaningful as it considers its utility and generalizability based on the conducted G studies.

2.8.3.6. The Consequential Aspect of Construct of Validity

Messick (1989) has introduced consequential validity as an essential component of argument-based approach to validity of tests. It "appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice, as well as to washback" (Messick, 1996, p.9-10). From the quote, it follows that Messick stresses the need to probe into the consequences of testing; evaluating actual and potential, short term and long term, intended and unintended consequences of score interpretation and utility. Issues of bias, fairness or unfairenenss, washback as sources threatening validity should also be investigated as they affect consequences and inferences of test scores.

In order to guide accurate test interpretation, Messick (1995) suggests that a researcher should provide accurate descriptions for the labeled test scores, specifically as a feature in "the consequential basis of validity" (p.745). In this sense, a test developer advocates the conventional approach to validity whereby s/he has to specify apparently what a test is intended to measure, determining the intended consequences of test use henceforth. Messick (1995, 1996) further emphasizes oversimplification in test score labelling to avoid ambiguous or rather misleading interpretations. For

example, ‘intelligence quotient’ is a construct that may make stakeholders think that is it a one-dimensional construct, whereas in reality the literature indicates that it is a multidimensional construct having multiple intelligences. Furthermore, he urges test developers to appraise the actual and potential consequences of test use.

In educational contexts, consequential validity evidence is related to the consequences (impact) that a test may have on policies concerning curriculum and instructional content, and school funding (Iliescu & Greiff, 2021). At this particular level, Messick (1989) believes that consequential validity should not be considered apart from other types of validity evidence. Still, it is the least used by test developers and users, and perhaps the most debatable concept amongst (Iliescu & Greiff, 2021). The authors assert that validity evidence is said to be an integral part in construct validity, because decisions made on validity information obtained from a given test emphasize whether the assessment represents the target content domain in regard to knowledge, attitudes, skills and behaviors, and evaluates how far the test consistently, fairly and authentically samples the intended construct neglecting most of the time consequential information. As such, tests developers and users focus on convergent information obtained from a respective test.

Consequential validity can be conceived of having positive or negative consequences. A good achievement test of productive vocabulary knowledge measures test takers lexical ability correctly, sampling the content area, and thus producing reliable scores on the intended competence of examinees under study. This test generates significant information leading to predict thoroughly learners’ outcomes, whether positive or negative such as, decisions to pass an exam, admission to higher educational levels, or to fail either repeating the year or school dropout. At a narrower level, consequential validity considers all what a test can generate as outcomes; however, further negative outcomes can also be predicted such as students’ demotivation to carry on a course or shift to joining other courses. As to teachers, test-driven education is an example of consequential validity; they might teach to the test based on test feedback instead of highlighting students’ ability to use vocabulary in real life settings for instance.

In Iliescu and Greiff's (2021) terms, if tests have to be judged not only at level of their ability to assess what they are intended to assess; their ability to generate good correlations with the intended achievement criteria, but also on their actual consequences in terms making decisions, then test users according to Iliescu and Greiff, must assume responsibility of actual and potential consequences a test may lead to. In brief, consequential validity digs up assessment procedures to better instructional practices enhancing learning and favoring teaching focused abilities.

In practice, the six aspects of validity evidence underlying the notion of construct validity, as a unitary approach underpinning validity theory, operate as standard criteria for all educational behavioral measurements, especially those connected with performance assessment (Messick, 1995).

Performance and competence assessments contributed enormously to the development of classical quality criteria of reliability and validity that were interestingly taken to estimate measurements. Reliability and validity were further enhanced in a sense that they are no longer traits of tests but they become traits of test scores and inferences. Additionally, when estimating reliability, several sources of variance of any measurement situation can be treated at once in a single analysis, with reverse to traditional measurements that used to tackle only one single facet at a time among the various measurements situation facets, including items, tasks, occasions, contexts, raters, teaching methods, etc.

Validity has become an inferential property or validity evidence that depends on complementarity and assimilation of arguments. These arguments have been classified under a unified concept or heading named construct validity evidence (Messick, 1995), within the same main stream, Messick emphasized predictive validity evidence as a crucial notion that holistically encompasses all types of validity advocated by American Psychological Association, Educational Research Association, and National Council on Measurement in Education (AERA, APA & NCME1999).

The aforementioned advancements were a consequence of that ambiguity associated with the concepts of reliability and validity from performance and competence perspectives. That is why it was highly demanded to operationalize them

in accordance with performance and competence tasks (Scallon 2004; Baartman et al., 2006; Baartman et al., 2007).

Overall, classical terms of reliability and validity increasingly gained great interest in performance/competence based assessments, but they still did not suffice the needs of judging the quality of these assessments. Consequently, Current trends in assessment have proposed many other recent distinctive quality criteria besides reliability and validity to judge the quality of performance/competence assessments (Linn et al., 1991; Linn, 1994; Johnson et al., 2009; Baartman et al., 2006; Baartman et al., 2007).

Up to this point, the six aspects of construct validity proposed by Messick have been discussed in some details, the subsequent section is devoted to a full description of modern quality criteria used to validate performnce/competence assessments.

2.9. New and More Quality Criteria of Performance Assessment

More recently, a new outlook in the role of assessment has been reflected in education. There was a transition towards assessment for learning at the expense of assessment of learning in the context of a learning culture (Shepard, 2000). For recall, traditional assessment centred on what students know in relation to curriculum outcomes. Thus, the assessment perspective revolved around an examination of validity and reliability issues of achievements tests in particular as the emphasis turned around assessment of learning whereby educationists consider knowledge and skills that learners gain as a result of instruction. A view of assessment for learning highlights both the end result or the final product, and the process of learning. Competencies assessment seems extremely intricate, mostly attributable to the fact that a competency consists of a multiplex integration of knowledge, skills and attitudes (Van Merriënboer et al., 2002, cited in Baartman et al., 2006). Since competence is complex in its nature its assessment, therefore, calls for different methods and standards.

The relationship between assessment and learning is evidential. This could be explained in the whashback or the backwash effect. Since instruction and learning are competency-geared, assessment should be aligned to competency-based principles. Traditionally speaking, reliability and validity were used as two means to measure achievement tests quality. In the present day, however, ten criteria have been proposed

to effectively assess the acquired competencies. Conventional assessment methods should not be casted off in CAPs because those measures of reliability or validity are not basically incorrect for CAPs, yet, they should be used differently and be regarded jointly with other quality criteria that are particularly essential for competency-based assessment practices (Baartman et al., 2006). In effect, a number of quality criteria have been suggested by different authors and educational researchers (Linn et al., 1991a; Linn, 1994; Baartman et al., 2006; Baartman et al., 2007; Johnson et al., 2009).

Some time ago, researchers as well as educational theorists and practitioners assumed that the focal criteria in achievement testing were, by all the means, reliability and validity as they contribute to the quality of assessment. Afterwards, new philosophy in assessment dominated the academic scene and directed the assessment practice from mechanistic choice to evaluation. This new tendency and new emphasis on new forms of assessment suggest that reliability and validity are of central interest but not enough or even not adequate to assess effectively and efficiently students' performance and competence. This persistently urged some scholars such as, Linn, Baker and Dunbar (1991,b), Baartman, Bastiaens, Kirschner and Van der Vleuten (2006), and Baartman, Prins, Kirschner and Van der Vleuten (2007), and Johnson, Penny and Gordon (2009), to develop new decisive and critical quality criteria for performance/competency assessment programs elaborating on conventional reliability and validity. Researchers' efforts effectively brought about practical and appropriate evaluative assessment designs.

Initially, Linn et al. (1991,b) introduce a comprehensive framework all-encompassing eight (08) quality criteria to improve the quality of evaluative evidence and support high quality performance/competence assessment systems (or CAPs). These criteria highlight consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, cost and efficiency. Additionally, Baartman, Bastiaens, Kirschner and Van der Vleuten (2006) and Baartman, Prins, Kirschner and Van der Vleuten (2007) elaborated a wheel of competency assessment composed of ten quality criteria to judge competences evaluations used in CAPs, derived from Linn, Baker and Dunbar (1991) assessment

framework. These include: authenticity, cognitive complexity, meaningfulness, transparency, educational consequences, fairness, directness, reproducibility of decisions, comparability, cost and efficiency. Formerly, Linn (1994) lists a number of other quality criteria linked to content, fairness, transfer, generalizability, comparability. At this point, and subsequently, Johnson et al. (2009) also propose eight useful quality criteria necessary to judge any kind of performance, a part from MC testing. These inclue: authenticity, context, cognitive complexity, in-depth coverage, examinee-structured response, credibility, cost and reform. We will briefly expose most of these quality criteria as they serve as foundational grounds in designing performance tasks, and in the validation process of the present productive vocabulary test.

In the following sections, a brief account of the quality criteria relevant and workable to the current study will be provided, and these in turn will help in taking measures as concerns designing tasks for the performance evaluation, investigating them in the different procedures taken in performance/competence evaluation.

2.9.1. Authenticity

A key feature in designing and assessing students' performance/competence. It relates to the extent to which assessment tasks align with the knowledge and skills considered critical to the domain of interest (Wiggins, 1998; Johnson et al., 2009). It implies the extent to which CAP resembles the future professional life (Gulikers et al., 2004; Baartman et al., 2006).

It is relatively a recent conception in the assessment literature. It is applicable to professional and/or everyday life situations those that are related to learners real world needs in particular. For example, using a phone call to arrange a meeting/appointment with a doctor, or writing a letter applying for a job are two tasks intended to elicit student's ability to solve real world problems using previously acquired knowledge, be it lexical knowledge. Authenticity means that if students are asked to produce a piece of writing, their writing skill should show how consistent this piece is with the real life needs of learners. Some reaserchers like Kuliker et al. (2004) declare that authenticity is varied according to five dimensions: the assessment task, the physical or natural

context, the social context, the assessment result or form, and the assessment or evaluation criteria.

Authenticity, therefore, refers to the degree to which a given assessment embodies knowledge and skills considered significant to a certain field. Authenticity refers to the knowledge and skills targeted in a particular assessment that should reflect or resemble those of a task accomplished in a workplace. For example, to hold a business meeting, a student enrolled in English for business courses, requires some knowledge of the language used particularly in the field related to how to do a business, and some skills for meeting management skills (e.g., turn taking, presenting a project using data shows). This kind of assessment would ask a student, for instance, to write a dialogue between a businessman and other businessmen sharing the same interest using some supposedly acquired knowledge and skills. As such, authentic performnce represent the knowledge and skills required to perform real world tasks found on a certain job.

The National Capital Language Resource Center (NCLRC, n.d.) suggest a set of well-defined criteria that any performance assessment activity must meet to achieve some degree of authenticity:

- be organized around interesting themes or issues to the learners;

- integrate real-world communication contexts, prompts and situations;

- ask students to produce an excellent product or performance;

- include multi-stage tasks and real world problems that encourage students to use language;

- make self-evaluation and self-correction possible when leaners progress in their learning; and

- evaluation criteria are known to examinees.

Overall, the importance of authenticity lies in judging the quality of performance/competence assessment. Authenticity, as a realistic property, denotes the extent to which fundamental knowledge and skills evaluated can be applied or reflected in the real or professional life. Authenticity, an aspect important in competency

assessment, is almost discarded in Messick's unitary framework of construct validity, but it will considered in the present test contruction.

2.9. 2. Cognitive Complexity

It entails that assessment tasks should incorporate high order thinking skills (Baartman et al., 2006). The authors equate cognitive complexity to authenticity due to the fact that both entail the use of thinking processes in the future professional life. The criterion quality of cognitive complexity depicts how well examinees' performance demonstrates the thinking processes used to solve problems in real life situations specifically related to future occupational settings. This quality is somehow linked to authenticity as it seeks to depict high order thinking skills, such as creative and critical thinking skill used by individuals to solve problems; to construct a response, apply knowledge and put simply, the degree to which the assessment reflects the existence of the cognitive skills and allows the assessment of thinking processes (Baartman et al., 2007).

2.9. 3. Context

The quality of context in assessment is determined by its authenticity. Authenticity promotes learning to take place and makes learning more meaningful and motivating as learners feel that there is a real purpose for learning, especially when performing tasks contextualized within the real world or natural settings. Performance tasks should be exposed in real-world contexts to allow students know that their skills and knowledge are valuable outside the classroom (Johnson et al., 2009). The researchers assert that the criterion quality of context is very critical in performance assessment because it "frames the design of tasks to assess complex skills within the real-world situations in which the skills will be applied" (p. 15).

2.9.4. Meaningfulness

Meaningfulness involves the degree to which competency assessment is valuable for all stakeholders including students, teachers, and employers (Messick, 1994). Baartman, et al. (2006) suggest three ways to make meaningfulness possible within the assessment tasks. First and foremost, by involving students in the development of the

assessment. Second, when their personal preferences are reflected in the assessment tasks so as learners can appreciate the assessment as being valuable; they comprehend the relationship between that assessment and their interests. Finally, an assessment is said to be meaningful when test takers take part in the assessment process, especially decisions made on when to take the assessment. That is, learners are not asked to produce until they feel that they are ready to produce and thus they can take maximum profit from the assessment. Assessment for learning is also relevant and meaningful when it rises learners' awareness of their points of strengths and weaknesses to stress further areas improvement (Schellekens et al., 2023).

It is worth mentioning that the quality criterion of "fitness for self-assessment" is subsumed by the quality of meaningfulness. Fitness for self-assessment means the extent to which CAP promotes the development of students' self-regulated learning, encouraging self and peer assessment, giving and receiving feedback (Baartman et al., 2013).

2.9.5. Fairness

Fairness is a factor effective in judging the validity aspect of an assessment system. It alludes to the degree to which examinees have an equal opportunity to show their performance/competence by means of reducing raters' bias (Baartman et al., 2006). A possible way to achieve fairness is to eliminate bias. Linn (1994) stressed that inequality of rating opportunities among members of minority groups, falsified interpretations of "between-group differences in average performance" (p. 569), and non-equal or unfair instructional chances are clear and direct causes of unfairness. Fairness implies that any assessment task should not display bias to minority groups of learners at the expense of others, and the knowledge, skills and attitudes of the competency in question should be discarding non-relevant components including content (Linn et al., 1991a). Other possible sources of bias are related to unsuitable accommodation to the learners' educational level or to tasks embedding features of culture that are not well-known for all learners (Baartman, et al., 2006).

2.9.6. Transparency

According to Oxford Languages (2023), Googles English Dictionary, transparency refers to the "quality of allowing light to pass through so that objects behind can be distinctly seen" or to "the quality of being easy to perceive or detect". With regard to performance/competence assessment, transparency associates with the degree to which the CAP is understandable and intelligible to all stakeholders, including students, teachers, and employers. It is thus that quality of being achieved in an open manner without covert practice. Assessment might become more transparent when leaners are informed of the scoring criteria, who are the judges of their performance, and what is the purpose of the assessment (Baartman et al., 2006). The authors further suggest inspecting if students are capable of evaluating themselves and their peers faultlessly as if they are trained raters. Consequently, peer and self-assessment present practically an index for transparency.

2.9.7. Educational Consequences

Messick points out to educational consequences and the effects of washback in his theoretical conception of consequential validity in his unified framework of construct validity (1994, 1995). Messick (1996) views washback as the effect of a given test on teaching and learning processes, which is one aspect of construct validity evidence that contributes to the validity of language test interpretation and use. For some authors like Linn et al. (1991b), Baartman et al. (2006), educational consequences is one quality criterion for competency assessment, which pertains the extent to which the CAP results in positive effects on teaching and learning, and the extent to which negative consequences are reduced.

The learning impact or educational impact as equated to educational consequences in the literature (Schellekens et al., 2023) refers to the intended and unintended, positive and negative effects of the assessment on the perceptions of both teachers and learners of education goals and accommodating the teaching and learning practices correspondingly (Linn et al., 1991a; Schellekens et al., 2023).

From the assessment for learning perspective, educational consequences tend to check whether assessment enhances learning. (Baartman et al., 2007; Schellekens et al.,

2023). Evidence of this impact lies in students learning via assessment especially when they develop an awareness of strengths and weaknesses (Schellekens et al., 2023) on the competency at stake. The effects of an assessment on student learning stresses on the interpretation of test scores and the learning impact related to it (Baartman et al., 2007). Based on test score interpretations, the assessment for learning culture uses scores to indicate student degree of mastery and prospective areas of improvement. Assessment of learning scores are used to display how well knowledge and skills on teaching are gained and teachers care much about teaching learners to succeed but not to promote learning (Schellekens et al., 2023).

2.9. 8. Transfer and Generalizability

Scores on a test seem to be indifferent in themselves and become more meaningful only if they result in valid generalizations about students' achievements (Linn et al., 1991b). Justifications for this quality criteria in assessing performance assessments lay in the application of G theory as a framework used for examining the level to which performance assessments scores are generalizable. For performance assessments to be generalized, a minimum of variability magnitude due to raters and tasks is necessary to be investigated in order to draw conclusions or generalizations (Linn et al., 1991b). For example, transfer or generalization from certain assessment tasks to larger universe of possible tasks of achievement have to be explained (Linn et al., 1991a).

Generalizability, therefore, relates to the dependability of assessment results (test scores) across facets of measurement such as raters, tasks, context and occasions (Linn, 1994). These facets of measurements should be regarded in the generalizability study designs and analyses to further indicate how far assessment results are dependable and generalizable. In these data collection procedures four specifications need to be regarded: rater training and drift, inter-rater agreement, inter-task generalizability, and generalizability across assessment types and contexts (Linn et a., 1991 b).

It is worthy to mention that learning transfer is another feature important to assessing the quality of performance/competence assessments. The learning transfer should be emphasized in performance tasks as they engage learners in new problem solving situations whereby learners demonstrate their ability to apply previously

acquired knowledge and skills to solve problems that they have never heard of or seen before. As such, the knowledge transfer is by no means a cognitive ability that allows test takers to transfer their knowledge or skills mastery from one situation to another. In class, a creative student might learn about the lexical field of cleanness: decay, bacteria, gingivitis, toothpaste, tooth brush, dessert, candy, treat, sweet, and write a paragraph about prevention of tooth decay, s/he can respond to a different assignment. In other contexts, the learning transfer occurs at level of task performance, where learners tend to transfer their knowledge and skills from one task to the next whereby they alter their strategies to respond effectively to a variety of tasks (Parkes, 2001; Parkes et al., 2000).

Under the criteria of generalizability and transfer, reproducibility of decisions can be discussed, because in the literature, it is a criterion related to Messick's generalizability aspect of construct validity. Messick (1994) has previously questioned if assessment results can be compared or even applied to other populations, contexts and tasks. Reproducibility has been defined as "the extent to which decisions made from the results of CAP are accurate and constant over situations and assessors" (Baartman et al., 2006, p.159). It is analogous to generalizability in the sense that if reproducibility increases generalizability also increases, it stresses on combining information sources rather than comparing different tests (Baartman et al. 2007). The authors draw a border line between reproducibility and generalizability, the former can be achieved by accumulating various information sources in a CAP (e.g., assessors, tasks, settings, occasions) in order to gain a clear and complete image of student's competences. The latter implies a comparison of different tests measuring the same construct to ensure generalizability. It is possible here to compare different tests because they yield a true score (from a true test) to which other tests can be contrasted. More precisely, when a researcher cannot reproduce the results of a study, approximately, any conclusions drawn from the original study are mistrust and generalizability is restricted (Downing, 2004).

In a nutshell, reproducibility is one principle of scientific objectivity that suggests that CAP should generate consistent and accurate conclusions and generalizability in

the long run. Note that in case of performance assessments where raters subjectively score students' performances, making decisions about the students is done accurately and do never be based on the judges or the context of the assessment. (Baartamn et al., 2006). This means that the results can be meaningfully reproduced and interpreted if the assessment data reached an acceptable reliably index (Downing, 2004).

2.9.9. Comparability

As its name suggests comparability of an assessment results refers to the degree to which the inferences drawn from the scores on one test can psychometrically be compared to the results of another test sharing the same assessment conditions. To achieve the criterion of comparability, "assessment tasks, criteria, working conditions and procedures should be consistent with respect to key features of interest" (Baartman et al., 2013, p.982). More precisely, a test should be carried out in a more consistent and responsible manner. It should be conducted under similar conditions, similar assignments should be administered to learners implementing consistently similar scoring criteria to all learners' performances. To increase comparability, Van der Vleuten and Schuwirth (2005) propose careful or rather random sampling among conditions of assessment in addition to utilizing large samples of content and situations selected for assessing the competency in play. Inferences based on biased sampling generate non-trustworthy conclusions and thus comparability is increasingly impossible.

2.9.10. Costs and Efficiency

The importance of cost effectiveness is evident, especially in the context of performance and competency assessment. Because CAPs are predominantly very intricate compared to classical tests and are, to a large extent, strenuous to conduct (Linn et al., 1991a). MC testing had arguably proved to be efficient and inexpensive even in high-stakes standardized assessments. This criterion quality assessment associates with the time and resources required to construct and conduct the CAP, compared to the profits. If extra time and resources are needed for further investment they should be justified by the positive consequences they might result in, being among teaching and learning improvements (Baartman, 2006). In order to determine if an assessment has

fulfilled cost efficiency, it is salient to consider all costs including resources and time invested with running the assessment and whether they have achieved high-quality. Cost efficiency means providing high quality assessment without maximising expenditures; saving money and time. One possible strategy to achieve this end is to pay special attention to the development and conduction of data collection procedures and scoring methods. CAP ability to achieve low cost, and in the same time improve assessment quality, and generate more profits compared with same time and resources and costs. Overall, assessment procedures should minimize expenses and maximize the benefits and values in the meanwhile.

2.9. 11. Fitness for Purpose

It relates to the level of alignments between assessment, standards, curriculum, and instruction (Baartman et al., 2006, 2013). Competency assessment should meet the principles of instruction, cover almost or all competencies, and assessment tasks should emphasise the integrated use of knowledge, it should assess knowledge, skills and attitudes (Baartman et al., 2013). Fitness for purpose also alludes to "the fulfilment of certain expectation" (Jung et al., 2016, p. 2). In performance/competence assessments, when constructing a test, content selection, and item sampling should meet defined instructional objectives.

2.9.12. Acceptability

It is a quality criterion critical to an evaluation of performance/competence measures because it determines its validity and effective use. The notion of acceptability means the degree to which students, teachers, employees and stakeholders are satisfied or consent to the assessment criteria and how CAP is conducted (Baartman et al., 2006, 2013). It is all about what attitudes stakeholders hold towards the fairness and acceptability of a rating system (Hedge & Teachout, 2000). Acceptability is deeply inherited in educational consequences or what is known as the backwash effect, an aspect of construct validity that has already been discussed in Messick's validity framework.

2.10. Reliability and Validity in Generalizability Theory Framework

According to Cronbach et al. (1963), validity and reliability are interrelated concepts; they can be classified under one heading in generalizability measures. The distinction, however, lies in the facets that we want to generalize on (Allem, 2000). On that account, generalizability theory is regarded among the recent measurement theories that gave birth to reliability and validity to be complementary concepts. In other words, it is possible to invest results obtained from reliability of behavioral measurements to conform validity of an assessment system, particularly, when it comes to construct validity. Through analysis of variance components and by limiting the universe of admissible conditions, a researcher can firmly explain variance of performance in tests, because ensuring validity of a given measurement procedure depends on a determination and description of different stimuli to which students react, such as instructions and rules to assess his responses.

Similarly, Quellmalz (1991) and Moss (1995) indicate that drawing a border line between reliability and validity in G theory has become blurred. There is a potential in performance assessment to take interaction effect between tasks as evidence for both reliability and validity. Therefore, the conventional distinction between these two notions becomes unclear and unnecessary in measuring the artistic characteristics of performance assessment (Mc Bee & Barnes, 1998). In the same mainstream, for Shavelson and Webb (2009), the leading figures in measuring performance, it is also possible to consider some facets of a measurement situation as facets of validity in G theory, and they might focus on different content areas and task type (multiple choice, open- ended, hand on tests,… etc.). These validity facets can be hidden in designing a G study.

G theory is a fully effective method used in the assessment of measurement accuracy. It can also serve as a method in the examination of construct validity of performance assessments via analysis and interpretation of variance components resulting from different facets of a measurement situation, namely persons, tasks, raters, occasions, assessment methods, contexts. From the perspectives of the recent validity theory that makes certain the complementarity of validity evidence under construct

validity, the conceptual framework advocated by Messick (1995) that includes generalizability as an important aspect, among other aspects of validity evidence, that emphasizes the potential of getting similar results across levels of different facets of the assessment process (e.g., tasks or raters). Huang (2012) on the other hand, views variance components resulting from different levels of facets central evidence to construct validity. The latter represents a useful method to investigate score validity in performance assessments. The more variance components of some facets increase, the more they are considered evidence for validity, and vice versa. That is why some facets of reliability of performance assessment scores such as tasks, occasions, and raters have been important evidence of validity.

Adversely, some opponents like Shavelson and Webb (2009) condemn reliability facets like tasks, occasions, raters, formats to be validity facets as different assessment methods, different testing content, different occasions. When we move from prototype measurement facets associated with test reliability to include facets linked to validity, this means that we shift from traditional reliability boundaries to adopt convergent validity in G theory. Shavelson et al. (1993) comment on the relationship of complementarity guiding reliability and validity in G theory framework in a performance assessment study through examining variability of treatments of different methods of task presentation convergent validity evidence applying the method-sampling variability across several methods of task presentation.

So far, the factors and processes of designing, developing and validating performance/competence assessments have been further explored, the subsequent section is devoted to developing appropriate rating scales to assess students' performances.

2.11. Scoring Performance/Competence-Based Assessments

In performance-based instructional programs, assessment plays a critical role in eliciting both what students know and what they can do. It does so by evaluating varying levels of students' performance through their responses. Within a culture that emphasizes *assessment for learning*, traditional one-to-one scoring relationships, where a correct answer earns a score of one and an incorrect answer earns zero, become

obsolete. This binary scoring model, which dominates classical testing formats such as multiple-choice and fill-in-the-blank questions, fails to capture the depth of student understanding or skill. In contrast, performance assessments task students with producing their own responses, thereby offering a more comprehensive measure of competency. For example, a performance assessment in vocabulary would require a student to construct a written piece using target words in meaningful contexts, as opposed to simply identifying correct definitions from a list of predetermined options.

The rise of alternative assessment methods emerged as a response to dissatisfaction with objective testing, which inadequately measured higher-order thinking skills and failed to reflect authentic learning contexts (McBee & Barens, 1998). These alternative approaches prioritize the assessment of cognitive processes and real-world tasks. However, they also introduce new challenges, particularly regarding how performances are graded. Unlike conventional tests, performance assessments lack fixed answer keys. This necessitates a shift toward expert judgement, which is inherently subjective (Clauser et al., 1995; Perlman, 2002). As Baartman et al. (2006) note, assessors must "abandon the idea that assessment is an exact science in which a true score can be found." (p. 156)

In foreign language (FL) classrooms, for example, inconsistencies often arise in the evaluation process. One rater may prioritize linguistic accuracy while another may emphasize rhetorical structure or content relevance, resulting in divergent scores for the same student performance. These rater biases undermine score reliability and call for a standardized approach. To address this, educators increasingly employ scoring rubrics, which aim to unify evaluative criteria and enhance inter-rater reliability, particularly in higher education, where student performance is diverse and nuanced. Scoring procedures in this context require making informed subjective judgments about the quality of students' work (Perlman, 2002), and rubrics are crucial in standardizing these judgments.

Researchers such as Ruiz-Primo and Shavelson (1996), Shavelson et al. (1998), and Perlman (2002) emphasize that performance assessments consist of two essential components: a *task* and corresponding *scoring criteria* or *rubrics*. The task may take

various forms, including extended written responses, oral presentations, portfolios, or collaborative projects, formats that stand in stark contrast to objective test items. These tasks require students to apply high-level thinking and demonstrate integrated knowledge and skills.

Rubrics are defined as rating tools that help teachers determine the level of student proficiency in performing a task or demonstrating conceptual knowledge. According to Brualdi (2002), a rubric is "a rating system by which teachers can determine at what level of proficiency a student is able to perform a task or display knowledge of a concept" (p. 4). Brookhart (1999) views rubrics as explicit scoring schemes designed by educators and researchers to guide the assessment of students' processes, behaviours, and products. Similarly, Popham (1997) and Perlman (2002) argue that rubrics provide structured and consistent frameworks for evaluating constructed responses—especially where traditional assessment falls short.

A well-designed scoring rubric should include three main elements: dimensions (criteria or traits), definitions and examples that describe performance expectations, and rating scales that indicate levels of performance (Perlman, 2002). Kan (2007) further explains that these dimensions are typically arranged vertically in a table, listing the traits of the performance, while levels of performance (e.g., excellent, good, average, poor) appear along the horizontal axis.

To illustrate, a test assessing *productive vocabulary knowledge* might include three criteria: *recognition* (understanding word meaning), *word formation* (ability to use affixes correctly), and *contextual use* (appropriately integrating the word in novel situations). A response rated at the "advanced" level would demonstrate in-depth mastery of all three attributes. If these are absent or underdeveloped, a "limited" or "weak" rating would be assigned. These descriptive scales help evaluators make fair, consistent judgments.

Brualdi (2002) recommends using numerical rating scales (e.g., 1–5) to determine whether students meet the standards set in each criterion. Brookhart (1999), however, cautions against using more than six levels, as increased scale complexity can reduce clarity and effectiveness.

Two main types of scoring rubrics exist: holistic and analytic. A holistic rubric provides a single score that reflects an overall impression of a student's performance. Castle (2018) states that holistic rubrics are suitable for *summative assessments*, particularly when a general performance rating is required. These rubrics are time-efficient and useful when the primary goal is a quick evaluation of a finished product (Kan, 2007).

In contrast, an analytic rubric breaks the performance into its component parts, assigning individual scores to each trait or criterion. As Castle (2018) notes, analytic rubrics offer detailed and targeted feedback, making them ideal for formative assessments that support ongoing learning. They are especially effective in diagnosing strengths and weaknesses across specific skills and are used to generate an aggregate performance score by summing the scores across all dimensions.

Importantly, no single rubric type is inherently superior; rather, the choice depends on the specific purpose of the assessment (Kan, 2007). Both analytic and holistic rubrics can serve formative and summative functions, as they provide qualitative descriptions of performance standards. However, before implementation, rubrics must be piloted and validated to ensure they effectively measure the intended traits and serve the educational objectives.

Scoring rubrics offer a number of pedagogical and psychometric benefits. They enable assessors to determine whether students have met particular criteria and offer targeted feedback to guide performance improvement (Moskal, 2019). Arter (2002) contends that rubrics help teachers monitor student progress in specific content or skill areas, allowing them to tailor instructional plans accordingly. Furthermore, rubrics enhance inter-rater reliability, improving the accuracy and consistency of performance evaluations (Smit & Birri, 2014). By promoting valid judgments and minimizing bias, rubrics contribute to equity in assessment (Perlman, 2002; Shumate et al., 2007).

Finally, rubrics help generate reliable estimates of student performance across different conditions, including task types, contexts, and evaluators (Johnson et al., 2009; Shavelson et al., 1992; Gipps, 1994). In doing so, they uphold the principles of fairness,

validity, and reliability within the competency-based assessment paradigm, ultimately supporting both student learning and instructional effectiveness.

Scoring rubrics are data collection procedures used to obtain students' performance scores. In order to investigate the reliability of the scores to be obtained, it is necessary to examine the relative impact of sources of variance on score dependability. This issue is elaborated along the following section.

2.12. Sources of Variability in Performance/Competence-Based Assessment

There exist a number of sources of variability, known as facets of measurement in the literature of G theory (for more information see Chapter3). These facets have a relative impact on the measurement precision, because they contribute, more or less, to measurement error and score variability. On the quality of performance assessment, stakeholders condemn that reliability of performance assessment procedures is limited due to sources of error variance. A review of literature and previous research in performance testing indicate four major sources of variability common to estimating the psychometric conditions of tests scores. These are tasks, raters, occasion, and assessment method (rating scales) in addition to the interaction effects between them; facet interrelationships such as intertask or interrater reliability. Nevertheless, in this arena, many studies differed in the number of facets and sources of variance incorporating the study designs fittingly to the intended research aims, though sophisticated studies emphasized almost all of the possible sources of variance in one measurement situation.

It is possible to explain sources of variation in a performance measurement situation referring to Shavelson et al. (1993) who modelled performance assessment within a sampling framework. The authors consider performance assessment score as a "sample of students' performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters and measurement methods" (p. 215). Task sampling variability due to "increasing sample size from the subject matter domain of interest". A task is representative of the subject-matter domain. Sampling variability due to occasion corresponds to classical test-retest reliability measure. Sampling variability due to raters corresponds to the possible individuals that can be trained and

thus participate in the scoring process. These three sources of variability were at traditional times, and even in modern times, proved to be components decreasing the reliability index of measurements.

Conclusion

This chapter progressed from deciphering the concept of performance assessment to its validation. It emphasized the features specific to performance assessment and the various quality criteria used for its evaluation. Besides the psychometric conditions of reliability and validity, performance assessment newly applied criteria for assessing the assessment were further discussed, more or less, in light of CAPs (Competency Assessment Programs), construct validity evidence and generalizability theory. This chapter ended up with a discussion of adequate scoring guides that would serve the current research purpose, namely assessing leaners' lexical competence, and then followed by a description of the major and common sources of variability researched in performance testing.

Having displayed some important diversified issues in vocabulary and performance/competence assessments valiadtaion in the two previous chapters, the following chapter will explore generalizability theory, taking that the relation between generalizability and performance assessment is overlapped. We, thus far, explored the complexities of language proficiency and its validation. Nevethless, the interpretation of assessment results dpend not only on task design and development but also on the reliability and validity of the inferences drawn from obtained test scores. Eventhough performance assessments give a clear image of students' lexical abilitites, they are subject to various sources of measurement error, which threaten consistency and accuracy of assessment data. To address this intricacy, generalizability theory, a statistical framework, is used to assess the dependability of behabvioural measurements across different conditions, including tasks, raters, occasions, methods and modes, etc.

CHAPTER THREE

GENERALIZABILITY THEORY

Introduction

Generalizability theory, or put simply G theory, is an annex to Classical Test theory (CTT). The limitations of CTT can be considered as a foundation for the emergence of G theory (Cardinet, et al., 2010). A researcher like Brennan (2010) asserted that G theory is still based on CTT, as it adopts a number of its basic concepts and in making use of CTT statistical procedures when conducting generalizability studies and analyses. The aim of this chapter is to delineate G theory fundamentals, evolution and conception. Since these derive from CTT, it is believed useful to discuss G theory in connection with CTT.

This chapter theoretically overviews generalizability studies (G studies) and decision studies (D studies), the two necessary steps in the investigation of conditions of measurement those applicable in planning, assessing, and establishing the dependability of measurement procedures.

The chapter also explains the organized structure of how well a researcher can be selective of the best model for conducting more effective G studies and D studies, implementing five mandatory steps. Beginning with observation design (data structure), going through phases of estimation design (facet level sampling), measurement design (study focus), design evaluation (generalizability coefficients) ending up with decision studies (optimization design). Additionally, data collection designs, including random, nested, and mixed designs, are explored. The chapter formally concludes with a review of previous studies researching vocabulary and language testing within the framework of G theory followed up with the contribution of G theory in the research paradigm.

3.1. History and Development of Generalizability Theory

In this section, evolution of G theory is discussed with reference to CTT, because as it is mentioned in the introduction, G theory has been developed in the context of CTT. As such, it is believed to be important to discuss G theory alongside the limitations of CTT, which, in turn, paved the way for G theory to arise.

3.1.1. Limitations of Classical Test Theory

Classical Test Theory (CTT) is crucial to understand before comparing its limitations to the advantages of Generalizability Theory (G theory). Known also as true score theory or classical reliability theory, CTT provides a foundation for traditional psychological assessments. Spearman (1907) introduced the theory in his article "Demonstration of formulae or true measurement of correlation," which outlined the core principles of CTT. However, the modern form of CTT emerged from the works of Gulliksen (1950), Magnusson (1967), and Lord and Novick (1968). CTT dominated psychometric research in educational and psychological measurement before Cronbach's G theory gained traction in the 1970s.

CTT was used extensively to assess the psychometric properties, such as reliability and validity, of scores obtained from various measurement tools like tests, questionnaires, and observation grids. The Spearman-Brown prophecy formula, introduced by Allen and Yen (1979), was one such method for estimating reliability after altering the number of test items. CTT is based on assumptions that have led to the development of G theory.

The core of CTT rests on five main assumptions (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 1992a, 1997, 2000; Laveault & Grégoire, 2002; Bertrand & Blais, 2004; De Gruijter & Van der Kamp, 2008). First, an individual's observed score (X) is the sum of his/her true score (T) and a random error (E). For example, if a student scores 14/20 on a test, his/her observed score is 14, but the true score differs, as measurement error always exists. Second, the expected observed score equals the true score, meaning that the measurement error is expected to be zero. Third, true scores and measurement errors are uncorrelated. Fourth, measurement errors across different tests are independent. Finally, no correlation exists between measurement errors in different tests and the true scores on others.

However, these assumptions have faced criticism, especially regarding measurement error. Critics argue that CTT oversimplifies error, considering it as a single undifferentiated source (Miller, 2010), whereas G theory accounts for multiple sources of error. CTT focuses on a single error of measurement (E), neglecting other

potential sources of variability. For example, CTT treats occasion error as the sole source of variability when using test-retest reliability, excluding other factors like raters or individual performance variations.

While CTT's approach to analyzing one source of error at a time may seem limited, it can be an advantage in certain contexts, offering simplicity. Nevertheless, as Brennan (2010) notes, focusing only on one-error source limits the theory's utility, making it less effective in fully addressing measurement reliability.

In this sense, CTT has weak postulation in comparison with other recent methods. According to Lord and Novick (1968), Allen and Yen (1979), and Bertrand and Blais (2004), CTT is a model of a true score inferred from puny hypothesis. Nevertheless, it managed to address issues of measurement precision and accuracy, so what approaches CTT developed to ensure reliability of behavioral measurements?

3.1.2. CTT and the Concept of Reliability

Classical psychometric research developed various methods or reliability indices for estimating reliability of measurements. These are listed by Bolus et al. (1982) along the following lines:

1-Test-retest reliability: Also known as repeatability, this method entails administering a test on two distinct occasion intervals and, thus, using correlations to confirm score reliability (stability). The method further tends to estimate the consistency of a measurement procedure in ranking test takers overtime. This reliability index suits norm-referenced objectives.

Test re-test reliability, explained above, is a key feature towards the achievement of reliability in CTT. In this theoretical framework, test length is analyzed by means of the well-known Spearman-Brown prophecy formula for reliability purposes. In this perspective, individuals participating in the study serve the object of measurement for purposes of ordering them along a continuum in norm-referenced testing. Despite this fact, a researcher like Llabre (1978) never considered persons as a facet in generalizability studies though it is a factor in Fisherian terms.

2- Administering parallel forms: Administering parallel forms of the same test is another method used in estimating reliability of measurements. It provides an estimate of the test means and standard deviations and then using the correlations to estimate the reliability of parallel forms to see the extent to which the two measurements yield equivalent results.

3- Internal consistency: It is often necessary to check how well individual items are consistent on an instrument and can yield similar results. Internal consistency is said to be high if inter-correlations have arguably proven to be high and the reverse does hold true for low inter-correlations. Provided this, low inter-correlations can result from measuring differing abilities and "hence distorting measurement of the desired trait" (Bolus et al., 1982, p. 247).

4- Inter-rater reliability: Another related reliability index involves rating subjective tests such as, paragraphs, essays, spoken records or oral presentations, and any other type of responses to open-ended questions in general in which determination of rater consistency among subject performance is crucial; to what extent the rater is consistent is his judgments of essays, for instance, in two different occasions. Correlations obtained among individual rater's scores on the same performance on two different occasions are used to estimate rater consistency. This reliability index together with the above reliability indices have been criticized for their functionality is restricted to estimating only one or two measurement errors at the maximum.

The reliability index in the CTT paradigm has been questioned in various contexts. As can be noticed, CTT implements several measuring procedures to estimate reliability of test scores. However, it does not allow the decision maker to control other intruded measurement errors that can affect the observed score when one of these methods is being conducted. These methods of investigating reliability suffer from ambiguity when coming to explain issues of validity and reliability of test scores and measurements in actual testing situation. This disadvantage of applying several methods to provide an estimate of reliability objectively reveals the theory's inability to control various errors that have the potential to affect a measurement procedure. CTT's indices of measuring reliability are not characterized by enough flexibility that fit all possible problems of reliability that appear in all cognitive testing (Crocker & Algina, 2006). Furthermore,

developing a variety of reliability indices to provide an estimate of test score stability reliably limits CTT strengths over other recurring test theories, and this does not go hand in hand with its theoretical definition (Webb et al., 2006).

Additionally, and as it has been discussed earlier in section 3.1, this classical paradigm assumes that a person's observed test score in a specific test is composed of a true score and one random undifferentiated error. This was the weakest point for CTT as it has a narrow and limited view of the sources of measurement error (Allem, 2000). Alternatively, the theory focuses on one single source of measurement error because using one of the aforementioned reliability indices, it is difficult to distinguish sources of variance that should be surveyed for the sake of reducing the measurement error. In this regard, it is not possible for the classical model to differentiate between the many sources of error. Think of when evaluating examinees' oral performance. It is possible, for instance, to estimate the measurement error related to raters due to severity using parallel ratings parameters. In the meanwhile, it is not possible to estimate errors related to other variables such as, persons (variations in assessment due to tiresome, demotivation, health status and psychological conditions, … etc.), tasks intended for evaluation (difference in degrees of difficulty, difference in from and content, etc.), and occasion considerations (e.g., morning, afternoon, beginning of week, end of week when the test is administered).

The limitations of CTT have along been discussed in conjunction with the advantages of G theory. The table below (Table 3.1), sums up and compares the differences between the principles of CTT and G theory in the research paradigm.

Classical test theory	**Generalizability theory**
-True score	-Universe of admissible observation's score
-One identifiable source of "error" variance	- Multiple sources of identifiable "error" variances.
-One-way ANOVA	- Factorial ANOVA
-"What if" optimizing assessment method: Spearman Brown	- "What if" optimizing assessment method: design study

Table 3.1: Differences between CTT and G Theory (MacIntyre et al., 2011, p.2)

On the basis of the aforementioned discussions, CTT seems to be rather feeble. As Allem (2000) argues, CTT is limited in its treatment of reliability and validity, as it fails to distinguish between multiple variance components. CTT can only provide estimates of total values assigned to different sources of measurement errors from a test administered under specific conditions that is why conventional methods of estimating reliability vary. For example, reliability in students' open-ended writing can be estimated across two or three occasions, but CTT treats occasion as a single factor, neglecting other factors like task or rater effects. This approach, while useful for certain measurements, does not control other variables such as, test timing, task, raters, or setting, nor does it account for the interaction effects between them. Therefore, CTT is seen as insufficient, and a more flexible theory like Generalizability Theory (G theory) is needed. As Cronbach (as cited in Cardinet et al., 1976) states, "a test is not reliable or unreliable; one can simply generalize, to different degrees, from one observed score to the multiple means of the different sets of possible observations." G theory excels by estimating variance components and minimizing errors, thus ensuring measurement consistency and accuracy, making it ideal for generalization purposes.

We conclude by explaining the most appropriate justifications for applying G theory instead of CTT. The latter is not intensively used in different situations, because as already discussed, it is unable to identify sources of variability in a measurement situation, and many sources of error such as tasks, raters and passages might intrude within a testing situation and cause measurement error (Kumazawa, 2009). Dissatisfaction with CTT's unreliability of assessment decisions led psychometricians to adopt a new conceptual framework named G theory that has been established against, and as a reaction to, the defects of CTT. Yet, it is also viewed as a continuum to the CTT because of its utility and investment of some practical statistical classical procedures (Brennan, 1992a).

G theory came up with new thoughts to complement some traditional principles of assessment, particularly those of classical ANOVA, a statistical formula, used to estimate reliability of behavioral measurements. The weaknesses of classical test score

theory, therefore, gave birth to Generalizability theory (G theory as named and abbreviated by Cronbach, Gleser, Nanda, and Rajaratnam (1972), and Shavelson and Webb (1991)) that has revolutionized evaluation paradigm, both in theory and practice. The classical ANOVA paved the way to the theory to flourish presenting a follow up to CTT. Kumazawa (2009), Brennan (2010) and Cardinet et al. (2010), among others, consider G theory as an extension to CTT, as it implements ANOVA procedures to estimate reliability and find out the multiple sources of error that cause the undifferentiated *E* in CTT. The development of G theory throughout time will be explored in brief in the upcoming section.

3.2. History of G Theory

The genesis of G theory can be traced as far back up as the early sixteenth of the twentieth century. It emerged in the onset of 1963 onwards when Cronbach and his coworkers published three articles entitled: "*Theory of generalizability: A liberalization of reliability theory"* (Cronbach, Rajaratnam, & Gleser, 1963), *"Generalizability of scores influenced by multiple sources of variance"* (Gleser, Cronbach & Rajaratnam, 1965), and "*The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles"* (Cronbach, Gleser, Nanda & Rajaratnam, 1972). The devised articles report univariate[5] G theory and are considered the first footprints of present day G theory.

But, officially and by no means, G theory principles are rooted in the collective work of Cronbach, Gleser, Nanda, and Rajaratnam (1972) in which multivariate[6] Generalizability was presented. Nevertheless, earlier insights were tackled about half a century afore, even before 1960s, in Fisher's (1925) works on coefficient designs in his *"Theory of statistical estimation"*. Despite this, G theory is not solely limited to testing the hypothesis used in experimental designs, rather it estimates the random effects variance components (sources of variability), the object of study among statisticians in the late 1940, because methods of ANOVA in G theory are not similarly implemented as to those of experimental designs. G theory differs from them in terminology and types of designs applied, as it is different in terms of orientation and trend (Allem, 2000).

Some researchers believe that the statistical model of ANOVA is the only efficient method to treat psychometric properties of test scores obtained from measurement procedures. During the forties, Cyril (1941) established one method, known as Hoyt's reliability coefficient, in search of estimating test internal consistency reliability using ANOVA. Some scholars considered Hoyt as a pioneer in estimating reliability via ANOVA. Brennan (2001) refuted this fact and pointed out that Burt (1936) was a pioneer who approached measurement problems through analysis of test scores obtained from an examination applying ANOVA. Furthermore, the study carried out by Finlayson (1951) on treatment of reliability of grades resulting from essay writing in terms of variance components; the fact of estimating variance components for the first time led to his description as the father of modern G theory. In the meantime, Ebel (1951) in his article entitled *"Estimation of the reliability of ratings"* conducted one-faceted crossed design and nested designs investigating two errors of variance; one considers rater main effects the other does not.

Developments of G theory enlarged in the fifties consisting of various contributions of some researchers establishing its solid grounds. To name but a few, Gulliksen (1950), Cronbach (1951), Pilliner (1952), Lindquist (1953)[7] and Burt (1955) put more emphasis on linking reliability to ANOVA method. Lord's (1955, 1957, 1959) articles considered standard errors of measurement and reliability investigated through the binominal error model. It is worthy to note the relative effect that Ebel's previous work (1951) had on Lord's with regard to rater and item main effects, both researchers initiated relative and absolute errors of measurement in G theory. In sum, CTT enormously contributed to the birth and evolution of G theory by means of implementing ANOVA procedures that largely smoothened the path to develop new concepts that, in turn, helped in the treatment of complicated measurement situations.

Improvements in G theory were not limited to this particular stage but they were practically pursued by studying variance components that helped researchers to study reliability of scores in various conditions of complex observation. In 1980 and 1987, a group of researchers contributed to G theory presenting a change in perspective, namely introducing the principle of symmetry of G theory by Cardinet and his associates

Tourneur and Allal (1976; 1981) in two co-authored essays published in *Educational Measurement* magazine. The authors suggest that any dimension of generalizability design can serve as an object of measurement besides persons. This conception has been improved in two relatively recent references (Bain & Pini, 1996, Cardinet et al., 2010).

Simultaneously, generalizability studies were not static; they rather were elaborated by two well-known researchers called Brennan and Kane (1977), who brought the generalizability coefficient that fits major decisions made in criterion-referenced tests, in addition to Kane (1982), who contributed to the extension of G theory by presenting a model that shows how G theory should be applied to estimate convergent validity of test scores. Along the same line, Crick and Brennan created GENOVA in 1983, a computer program designed for conducting G theory analysis. Afterwards, particularly in 1989, Feldt and Brennan dedicated about a whole chapter discussing in details issues associated with reliability in G theory.

In the outset of the 1990s, more case studies have been conducted by Shavelson and his associates (e, g., Shavelson, Baxter & Gao, 1993; Ruiz-Primo, Baxter & Gao, 1993; Baxter, Shavelson, Herman, Brown & Valadez, 1993; etc…) applying G theory to investigate performance assessments in the fields of mathematics, natural sciences, languages and educational research. These studies strikingly received a set of interest with the publication of a primer (or a monograph) by Shavelson and Webb (1991) on G theory, the aim of which was to make the theory and its methods accessible to anyone wishing to explore the field. Onwards, studies applying G theory had enormously increased in performance based-assessment endeavor.

More recently, many researchers and authors carried on the task of developing G theory especially within the context of performance-based assessment. Brennan (2001), for instance, developed *Multivariate Generalizability Theory* in a reference rapidly spread among measurement specialists, in addition to other relatively recent reference proposed by Cardinet et al. (2010) who provided a detailed account of G theory and its applications using EduG software package. Other studies (e.g., Brennan, 2000, 1997; Huang, 2009; Alkharusi, 2012) emphasized different perspectives of measurement situation by estimating variance components.

Thus far, we have been concerned with the development of G theory along a string of advancements in the research endeavors. The next section is devoted to conceptualize the theory under study together with its advantages.

3.3. Definition and Merits of Generalizability Theory

Generalizability Theory (G-theory) has emerged as a pivotal framework in psychometric research, particularly in the social, natural, and health sciences. At its inception, Lee Cronbach, alongside collaborators like Nageswari Rajaratnam, Goldine Gleser, and Ralph Nanda, introduced G-theory as an extension of Classical Test Theory (CTT) to better assess the dependability of behavioral measurements (Cronbach et al., 1972; Cronbach, Rajaratnam, & Gleser, 1963).

One of the primary strengths of G-theory is its ability to simultaneously examine multiple sources of measurement error. Unlike CTT, which typically considers a single source of error, G-theory employs a statistical framework that decomposes observed score variance into components attributable to various facets such as person, item, occasion, and rater (Shavelson & Webb, 1991; Brennan, 2010). This approach allows for a more nuanced understanding of measurement reliability.

Furthermore, G-theory facilitates the identification of interactions between these facets, such as person-by-task or person-by-rater interactions, which can significantly influence measurement outcomes (De Gruijter & Van Der Kamp, 2005). By quantifying the impact of these interactions, researchers can make more informed decisions about assessment design and interpretation.

In addition to its analytical capabilities, G-theory offers practical utility in assessment research. It enables researchers to evaluate various reliability indices concurrently, including test-retest reliability, internal-consistency reliability, and inter-rater reliability (Yin & Shavelson, 2008). This comprehensive evaluation aids in optimizing assessment procedures to enhance measurement dependability.

While CTT has been foundational in measurement theory, it has limitations in addressing the complexities of modern assessment contexts. CTT typically assumes that measurement error is random and does not account for multiple sources of error

simultaneously. In contrast, G-theory provides a more robust framework by acknowledging and quantifying various sources of error, thereby offering a clearer picture of measurement reliability (Brennan, 2010).

In summary, Generalizability Theory represents a significant advancement in the field of psychometrics. Its ability to decompose observed score variance into multiple components and identify interactions between measurement facets provides researchers with a more detailed and accurate assessment of measurement reliability. By offering both theoretical insights and practical tools, G-theory enhances the rigor and validity of behavioral measurements across diverse research domains.

Generalizability Theory (G-theory) provides a robust framework for assessing the dependability of behavioral measurements by systematically analyzing the variance components attributable to various facets of the measurement process. This approach is particularly valuable in educational and psychological research, where multiple factors can influence measurement outcomes. A key feature of G-theory is its two-stage analysis process: the Generalizability (G) study and the Decision (D) study, which will be described in detail in the next section.

3.4. Generalizability and Decision Studies

In the realm of assessment research, Meyer (2010) delineates two distinct stages within Generalizability Theory (G-theory): the Generalizability (G) study and the Decision (D) study. The G study aims to identify significant characteristics of a measurement procedure and investigate the variability of scores associated with each characteristic, thereby estimating the sources of variance inherent in the measurement process. Conversely, the D study utilizes the data gathered from the G study to evaluate the dependability of measurement scores and to inform the planning of more effective measurement procedures.

A pivotal distinction between these two types of generalizability analyses lies in their objectives and methodologies. Data from G studies are employed to estimate the various components of variance obtained from specific methods, whereas D studies focus on utilizing this information to make informed decisions regarding measurement procedures.

For instance, in a G study, researchers collect data from a sample of participants observed across various performance tasks. The study design might be formulated as person-by-task (p×t) or, if raters are involved, as person-by-task-by-rater (p×t×r). Analysis of variance (ANOVA) is then used to estimate the variance components by calculating the mean squares derived from and within subjects. The G coefficient is subsequently calculated to determine the proportion of the expected observed score variance attributable to the best estimate of the true score (Bottema-Beutel et al., 2014).

In contrast, a D study employs the data from a G study to assess the dependability of scores under various conditions. It identifies the optimal method to achieve an acceptable degree of reliability, thereby enabling decision-makers to draw accurate generalizations of scores to the universe score. This process allows researchers to utilize information about the relative error contributions to optimize measurement precision for future research conducted under similar conditions (Cardinet et al., 2010; Bottema-Beutel et al., 2014).

For example, in a three-faceted design involving rater, task, and occasion as facets, a D study might explore how varying the number of raters, tasks, or occasions impacts the generalizability coefficient. Researchers aim to determine the best methodological approach that is both feasible and effective for broader generalization. The universe of generalization in D studies often differs from the universe of admissible observations, with admissible authentic communicative tasks being randomly selected from a universe of classroom various authentic communicative types of tasks (Brennan, 2001; Cardinet et al., 2010).

In the literature of G theory, conducting a D study is an attempt to determine whether increasing or decreasing variability due to facet levels and measurement conditions will increase or decrease reliability, respectively (Cardinet et al., 2010). If the G coefficient obtained is, for example, 0.80, researchers might wish to reduce the number of tasks from two to one and compare whether this decision will, in turn, increase or decrease reliability coefficients. The same holds true for reducing or extending the number of raters involved in the universe of generalization and the effects they may produce in the overall D study design. This process helps researchers adopt

these two approaches in future research.

Compared to G studies, D studies identify the universe of generalization rather than the universe of admissible observations. They probe into the extent to which results obtained can be generalized to the universe scores. D studies further allow researchers or decision-makers to design measurement procedures seeking to minimize errors and thus increase the reliability or dependability of measurements.

Drawing a distinction between G studies and D studies does not imply that they are entirely separate. In fact, the best experimental design on which a G study is built depends on the decision that the researcher seeks to construct (Allem, 2000). While some scholars view the D study stage as beginning with an explanation of generalizability coefficients obtained from the G study stage for further exploitation in future studies, others consider D studies to be distinct from G studies as they depart from measures used to broaden measurement procedures relying on estimating components of score variance (Marcoulides & Kyriakides, 2010).

Despite these differing perspectives, the distinction is significant, especially in studies with various viewpoints, because plans used in G studies differ from those used in D studies. Meyer (2010) notes that a researcher conducting a G study is concerned with plans that permit estimation of variability components affecting measurement accuracy. In contrast, when conducting a D study, the researcher considers plans that would help improve measurement procedures based on decisions made in the G study results and estimating random error used in the study. However, it is possible to exploit D study results to plan new G studies based on decisions made by a researcher in the D study phase.

Shavelson and Webb (2009) compare a G study to a pilot study in that it is planned to "isolate and estimate as many facets of measurement error in the universe of admissible observations as is reasonably and economically feasible" (p.15). The G study stage aims to identify the sources of variability in behavioral measurements. As for a D study, it can be seen as the practical part devoted to implementing information described in the G study to design a specific measurement instrument.

Three steps are essential when conducting a D study:

1. Defining a universe of generalization, which includes the facets and their levels over which decisions are made to generalize;

2. Specifying the proposed interpretation of the measurement (relative or absolute); and

3. Utilizing information obtained from the G study regarding sources of error variance to evaluate the efficacy of alternative designs for purposes of reliability achievement and error reduction (Shavelson & Webb, 2009, p.15).

These steps have been clearly explained, and the last step in decision-making is conventionally equated with CTT's Spearman-Brown prophecy formula. It explains the principle of cost efficiency in CTT, which means increasing the number of tasks will decrease error, thereby increasing reliability. It is further worth mentioning that CTT does not conduct G studies planned to estimate the maximum sources of variance as it emphasizes only one source of error at a time. Practical application of G and D studies is set forward by applying G theory.

3.5. Applying Generalizability Theory: Concepts and Principles

This section sketches out how a researcher can successfully conduct an educational or psychological measurement using G theory. Thus, its different terminology and phases at play should be identified along with its relevant mathematical equations (content). We rely on the steps set up forth by researchers like, Bain and Pini (1996), Laveault and Grégoire (2002), Bertrand and Blais (2004, based on Cardinet & Tourneur, 1985), and Cardinet et al. (2010) to clarify major concepts and principled procedures that are used in G theory studies. That is, to evaluate the properties of a given measurement tool and estimate measurement precision across different measurement situations. In their proposal, five basic stages are determined; the first two named observation and estimation designs are respectively associated with analysis of variance, whereas the third, termed measurement design, is associated with the G study stage and the fourth, called optimization, is related to the D study stage. These procedural steps and their related concepts are further explored in the upcoming sections in their order of appearance. Borrowing the terms naming the phases of G studies from Cardinet et al. (2010), this section will proceed to provide a comprehensive practical framework of G studies and D studies, the core of the next concern. These phases will

be implemented as methodological procedures used for data collection and analysis and optimization decision.

3.5.1. Observation Design

Observation design, also known as "observed universe design" or "data structure," is the foundational step in a Generalizability (G) study. It involves identifying the facets, such as persons, tasks, raters, and occasions, that constitute the measurement procedure and their interrelationships (Meyer, 2010).

The term "facet" was introduced by Guttman (as cited in Cardinet et al., 2010) to avoid confusion with factors in factor analysis. In G theory, a facet is defined as a set of measurement conditions (Crocker & Algina, 2008). For example, in assessing student writing, facets might include raters and rating designs like holistic or analytic rubrics.

Facets are considered sources of measurement error, and their interactions can impact the reliability of measurements. Shavelson and Webb (2009) describe a facet as "one major source of variation" (p. 2). A universe in G theory comprises multiple facets; for instance, a complex universe might include persons, tasks, raters, and occasions (Shavelson & Webb, 1991).

The universe of admissible observations refers to all possible observations that decision-makers consider acceptable substitutes for the sample observation at hand (Shavelson & Webb, 2009). This concept underscores the importance of defining the universe appropriately to ensure valid generalizations.

Briesch et al. (2014) identified several relevant facets in G studies, including rater, form, item, occasion, setting, method, and dimension. These facets can be combined in various ways to examine their interactions and their effects on measurement reliability.

Observation designs can be simple or complex, depending on the number and type of facets involved. Facets can be either crossed or nested. In a crossed design, every level of one facet is combined with every level of another facet. For example, in a person-by-item (p×i) design, each person responds to every item. In a nested design, levels of one facet are nested within levels of another facet.

To represent variance decomposition in observation designs, Euler-Venn

diagrams are often used. These diagrams, introduced by Cronbach et al. (1972), visually depict the sources of variance and their interactions in a measurement situation. They are valuable tools for understanding the complexity of measurement procedures and for designing studies that accurately assess reliability (Cardinet et al., 2010; Meyer, 2010).Possibly, the best procedure used to represent the variance decomposition is *Euler Venn* variance partition diagram, a set of intersecting circles or ellipses, initially put forward by Cronbach et al. (1972) to determine sources of variance in observed designs. *Euler Venn* diagrams are deployed by various researchers (e.g. Shavelson & Webb, 1991; Cardinet et al., 2010; Meyer, 2010) to describe measurement situations. *Euler Venn* diagrams reflect the measurement situation in G studies in terms of specifying facet number, number of observed levels, as well as the inter-relationships existing between every couple of facets (Bertrand & Blais, 2004). To illustrate, the two facet-crossed observation design (p×t or p×i) containing variance division is envisaged through the following Euler Venn variance partition diagram:

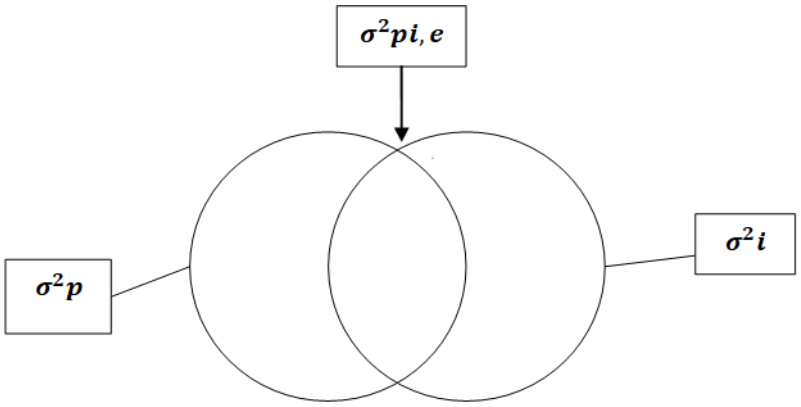


Figure 3.1: Variance Partition Diagram for the Two-faceted (p × i) Design (Cardinet et al., 2010)

Figure 3.1 provides the variance partition diagram for the simplest observation design **p×t** discussed above, where the two ellipses represent the three effects of students **(p)**, tasks **(t)** that are both random facets, and student-task interaction effects **pi**. The above partition describes the contribution of the three effects of students, tasks and student-task interaction to total score variance. The area of overlapping closed curves represents student-task interaction effect associated with **(e)** which is, the

variance resulting from random variation (or fluctuation) and unidentified sources of systematic variance (Meyer, 2010). That is, the variance of students observed scores due to variance components and unmeasured systematic sources of error (residual).

In other more complex crossed designs where measurement situation involves more than two facets, be it occasion, test items and tasks, and students that are object of measurement can also be considered facets symmetrically. In this example occasion facet; moments when a test is administered, let's say on day 1 and day 2 where students are required to respond to all test items, then the three facets of students, tasks and occasion are crossed. This observed design is denoted as **(p×t×o)** or simply **(pto)** and the crossing relationship is visualized through the following Euler Venn diagram (Figure 3.2):

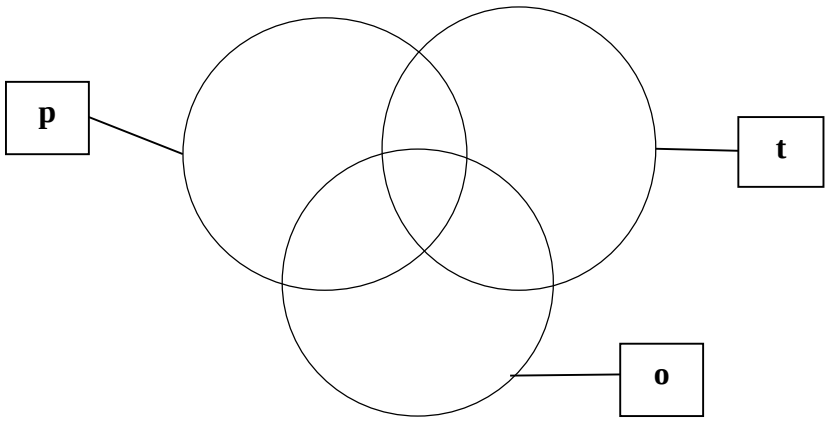


Figure 3.2: **Variance Partition Diagram for the Three Facet (p×t×o) Design (Cardinet et al., 2010)**

The observation design in Figure 3.2 displays that students **(p)** are crossed with occasion **(o)** (i.e., Day of administration) and tasks **(t)**. In this particular p×t×o three facet crossed design, task and occasion are considered random facets.

In other contexts, and as stated earlier, there are basically two types of facet interrelationships involved in G studies. Besides crossed-facet designs, nested designs are brought to existence. Two facets are described as nested "if each level of one is associated with one and only one level of the other" (Cardinet et al., 2010, p.13). Supposedly, there were 100 students who performed in a test of writing ability, whose performance is judged differently for comparative study reasons by two raters, each

scoring half a sample (50 students) or evaluating all 100 students but in this case using different rating schemes. In this regard, students are nested within raters and items are nested within scoring designs. None of these designs adheres to the same measurement conditions. The symbol set for the nesting relationship is signaled by a colon **(:)** and nesting students in tasks and raters is **(p:t:r)** read as "persons nested within tasks and tasks nested within rater". Nesting is like crossing relationships in that they might involve more than one measurement facet. The nesting facet relationship might be applied to students responding to diversified test items too. The same *Euler Venn* diagrams can visualize facets with nested designs such as **(p:t:r)** shown in Figure 3.3.

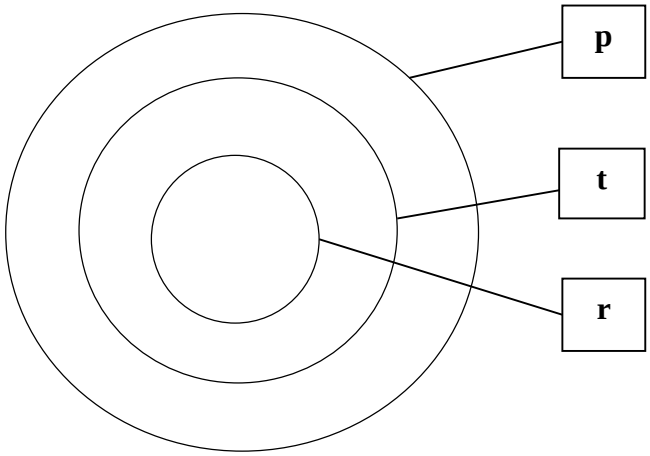


Figure 3.3: Variance Partition Diagram for the Fully-nested (p:t:r) Design (Cardinet et al., 2010)

In a similar vein, in case the observation design has two facets, where students and occasion are nested (symbolized as **p:o**) and simply **(p:o)** design would be read as "persons nested within occasion" using G theory terms. It displays a fully nested design with two facet components where participants are administered a test in two distinct occasions, let us say morning and afternoon, for example. When test items are scored by different raters the nested partitioning would be denoted as "students: raters" or simply put **(p:r)** and the same does hold true for "item: rater" **(i:r)**. This kind of observation design representing crossed-nested facet interrelationship is further diagrammed in Figure 3.4 as under:

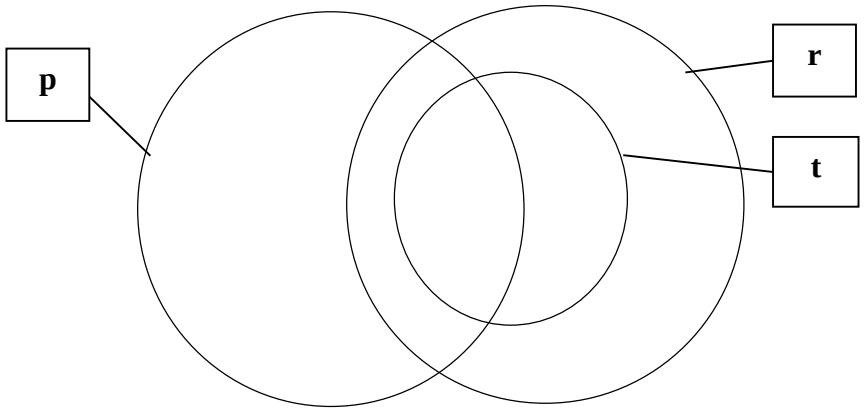


Figure 3.4: Variance Partition Diagram for the Partially-nested p×(t:r) Design (Cardinet et al., 2010)

It can be noted, here, that a facet may also be crossed with or nested within the object of measurement as it can serve as a constituent within a partially-nested design where, for instance, one facet is crossed with two nested facets. The above variance partition diagram designates how a generalizability design might be partially nested and, thus crossed with another facet, where raters and tasks are nested and crossed with students (or one person facet is crossed with two nested facets of task and rater). In this sense, this design is symbolized as **p×(t:r)**. It denotes that a group of learners belonging to different classes might be rated by a set of raters. One advantage of the nested facet design is its efficiency and economic effects (Cardinet et al., 2010).

In a nutshell, designs with crossed facets and partially nested facets allow us to decide about the variance components to be estimated in the current G study, where all students are observed under all facets and their conditions (levels and items). Several components will be investigated including person variance, item (task) variance, context (theme) variance, rater variance, random error variance, person-by-task interaction variance, person-by-rater interaction variance, item-by-theme variance, and a residual caused by the person-by-task-by-rater interaction variance, …etc. Accordingly, crossed designs are preferred in the G study in order to estimate all possible sources of variation in a student's observed test score where all participants share similar assessment conditions, an advantage which is totally lucking in the fully nested designs. The present study designs, however, will also incorporate partially nested designs justified by logistics (See chapter five).

In the observation design phase data structure is described from the perspectives of facets and their inter-relationships. But, for Cardinet et al. (2010) this phase is still incomplete because knowing the data structure associated with every observation design is not satisfactory neither to correctly estimate variance components nor to determine the basic G theory model, especially that facets have another trait to consider when conducting a G study, that of sampling status.

3.5.2. Estimation Design (Facet-Level Sampling)

Generalizability (G) theory applies a statistical sampling framework to assess measurement procedures, particularly through estimation design or sampling of facet levels (Cardinet et al., 2010). The way facets are sampled, whether fixed, random, or finite random, significantly affects measurement error (Shavelson & Webb, 1991; Cardinet et al., 2010; Bain & Pini, 1996). A facet is considered fixed when all possible levels are included in the dataset, such as when all grammar teaching methods are observed (Bertrand & Blais, 2004; Webb et al., 2006; SSREWG, 2010). In these cases, generalization beyond the sampled levels is not intended. According to Webb et al. (2006) and Shavelson and Webb (2009), fixed facets arise when (a) generalization is not the research aim, (b) it is illogical to generalize, or (c) the entire universe of conditions is already represented.

In contrast, random sampling involves selecting facet levels at random from the target population (Cardinet et al., 2010). This could involve sampling a single teacher training school from several, or selecting students, communicative tasks, or vocabulary items from a larger pool. Cardinet et al. (2010) distinguish between finite random sampling, where the population is limited (e.g., schools in Algiers), and infinite random sampling, where the population is vast or uncountable, as in the case of test items or students (Bain & Pini, 2006).

These distinctions between fixed and random facets have implications for score reliability and the extent of generalization (Meyer, 2010). Notably, Cardinet et al. (2010) emphasize that a facet's sampling status can be redefined after data analysis, depending on the research purpose. For example, if test items were initially treated as fixed but broader generalization is required, they may be reconsidered as finite random.

Finally, graphical representations (e.g., Figure 3.5) can clarify estimation design and sampling types. In a design such as p × o × t, fixed facets like ‘occasion’ are typically shown with dotted lines, while random facets such as ‘person’ and ‘task’ are illustrated with solid lines, using Venn diagrams to visualize variance contributions and relationships among facets.

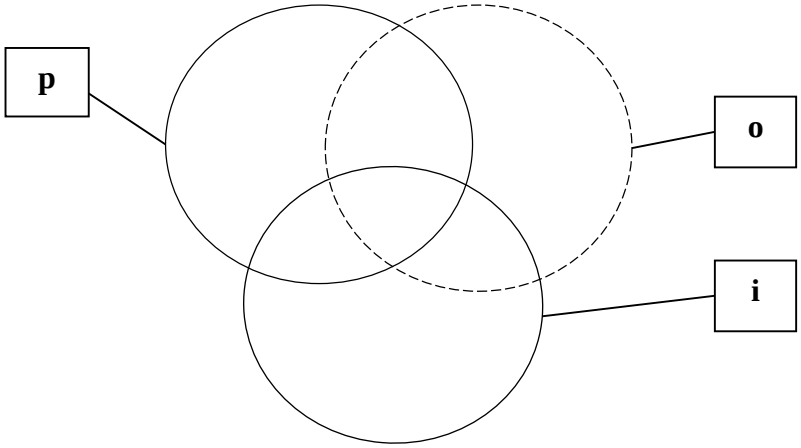


Figure 3.5: **Variance Partitioning Diagram for p×i×o with Occasion as a Fixed Facet (Cardinet et al., 2010)**

Inherent in the estimation design phase is another sampling status named mixed sampling. In a measurement procedure, mixed sampling can concurrently encompass both fixed and random facets and random samples can be derived from a finite universe (Cardinet et al., 2010). Mixed designs are well-discussed through section 3.6.2.

May be the best description set for random and fixed facets has been summarized by Shavelson and Webb (1991). Samples, or facets, are treated to be random when the magnitude of the universe is broader than that of the sample. The random model is mostly applicable in G theory paradigm being based on a random sampling theory. When conducting a random effects model, researchers should randomly sample the conditions (facets), and the universe of conditions should extremely be large. By contrast, a fixed model or a fixed facet that is generally equated with the traditional term fixed factor in CTT ANOVA refers to “the number of conditions of the subtest fact equals the number of conditions in the universe of generalization” (Shavelson & Webb, 1991, p.12).

It is worth noting that treatments of facet as being fixed or random will obviously

influence the generalizability of measurements. When facets sampling status is considered random, the universe of generalization would be large and for fixed facets, however, generalization is limited to the sample itself. Simultaneously, generalization in mixed sampling is, by no means, relevant to the kinds of facets and their conditions within a finite or infinite universe.

In the aforementioned stages of G study the different steps implemented in psychometric measurements are further identified; the facets of a measurement situation and their inter-relationships are determined in the data set for analysis purposes, and so thus the sampling status, whether fixed or random. These kinds of information are part and parcel in ANOVA used to estimate reliability and accuracy of measurements; they help in estimating variance components correctly. However, generalizability analysis cannot be completed unless supplemented by specifications of study focus.

3.5. 3. Measurement Design (Study Focus)

In the initial stage of a generalizability (G) study, researchers determine the object of measurement and decide whether the approach will be norm-referenced or criterion-referenced (Cardinet et al., 2010). This foundational step, often called estimation precision, identifies the key facets in the dataset and shapes how the data will be interpreted. In my study, students are the object of measurement, and both relative (norm-referenced) and absolute (criterion-referenced) approaches are applied to evaluate vocabulary performance.

Norm-referenced assessments compare students' performance against that of their peers, placing individuals along a continuum (Meyer, 2010). Criterion-referenced assessments, on the other hand, evaluate students against a fixed standard (Cardinet et al., 2010). G theory offers a flexible framework to support both forms of decision-making, depending on the research objective. The distinction is fundamental: norm-referenced (relative) decisions are useful when ranking students, for example, in competitive placement contexts, whereas criterion-referenced (absolute) decisions are more relevant in licensing exams, where meeting a minimum proficiency level is the goal (Shavelson & Webb, 1991; Meyer, 2010).

Relative decisions are tied to δ-type error variance and focus on ranking students

in relation to one another. In this case, individual scores are interpreted through statistical tools like the mean, standard deviation, or percentile ranks (Marcoulides & Kyriakides, 2010). For instance, selecting candidates for an ESP teaching program based on their writing scores involves identifying the highest achievers and comparing them to peers. Here, only the sources of error that affect relative standing matter (Shavelson & Webb, 1991). In contrast, absolute decisions are tied to Δ-type error variance and involve evaluating individual achievement levels against a benchmark. An example would be using a placement test to determine whether students meet the entry standard for a specialized teacher-training course. In such cases, students are treated as individuals rather than members of a comparative group.

Beyond these decision types, G theory extends classical test theory (CTT) by recognizing that observed scores are influenced by multiple sources of variance. Its purpose is not just to measure but to differentiate between students reliably. To this end, G theory focuses on three types of variance: differentiation, instrumentation, and generalization.

Differentiation variance refers to the true differences between subjects being measured. For example, in a (p × t × r) design, where students (p) are the object of measurement, the variance between students reflects differentiation variance. This component reveals how effectively the assessment distinguishes between learners (Cardinet et al., 2010). High differentiation variance indicates a more reliable measurement for both relative and absolute decisions.

Instrumentation variance arises from the conditions under which measurements occur. These facets, like tasks, themes, or raters, influence student performance differently depending on difficulty, familiarity, or complexity (Meyer, 2010). For example, task complexity or rater subjectivity could affect students' outcomes and thus contribute to measurement error. Instrumentation variance highlights the need to control or account for variability in task and rater conditions.

Generalization variance, finally, refers to random or residual errors introduced by the sampling of facet levels or unidentified sources (Cardinet et al., 2010; Colman, 2015). It encompasses the variance that hinders the generalizability of results beyond

the observed dataset. Lower generalization variance suggests more precise and stable measurements, essential for drawing broader conclusions.

Altogether, G theory enhances the measurement design process by allowing researchers to model and evaluate the impact of various facets, guiding better decisions in both norm-referenced and criterion-referenced contexts.

Estimating reliability of a measure necessitates computation of generalizability coefficients that are to be determined in the fourth stage called design evaluation.

3.5. 4. Design Evaluation (G coefficients)

Generalizability theory (G theory) addresses key measurement issues, particularly the estimation of variance components and the calculation of generalizability coefficients (G coefficients). The goal is to quantify the consistency of a measurement procedure, such as a test, in assessing the intended trait or achievement.

In contrast to classical test theory (CTT) [5], where the reliability coefficient is the ratio of true score variance to total observed score variance:

$$\boldsymbol{\rho^2_{XT} = \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(E)}}$$

Miller (2010) explains that the reliability coefficient ($\rho_{XX'}$) is the ratio of true score variance to observed score variance, which includes both true score and error variance. In G theory, the G coefficient is the ratio of universe score variance to observed score variance, with estimates varying based on the generalizability study design. G theory considers an individual's score as part of a universe score, influenced by multiple measurement conditions (e.g., raters, occasions), unlike CTT, which assumes a single true score for each individual under one measurement condition.

For recall, CTT reliability estimates are derived from three methods: test-retest reliability (Marcoulides & Kyriakides, 2010), internal reliability (split-half procedure), and parallel forms reliability (Yelboğa & Tavşancil, 2010). Additionally, inter-rater agreement (Yelboğa & Tavşancil, 2010) checks for consistency across raters. However, CTT's treatment of error as a single source fails to distinguish between different error types (e.g., form, item, occasion, rater), a limitation addressed by G theory (Shavelson

& Webb, 1991; Brennan, 2001, 2010).

In classical test theory (CTT), any observed score (X) is broken into two components: the true score (T) and random error (E), or 2X= 2T + 2E. This model simplifies reliability estimation compared to the more complex framework of generalizability theory (G theory), which accounts for multiple sources of error and varying measurement conditions (Shavelson & Webb, 1991). In G theory, the true score is replaced with the universe score, representing the expected score across specific conditions (Laveault & Grégoire, 2002).

The generalizability coefficient or G coefficient is thus defined as the proportion of universe score variance relative to observed score variance. Shavelson and Webb (1991) stress that the "dependability of measurement" lies in how accurately we can generalize from specific measurement facets, such as items, raters, occasions, to the broader universe. To ensure precision, G studies incorporate both facet variances and their interaction effects.

In CTT, reliability is the ratio of true score variance to observed score variance:

$$\rho XX' = \sigma T2/\sigma X2 = \sigma T2/(\sigma T2 + \sigma E2)$$

This coefficient ranges from 0 (pure error) to 1 (perfect reliability), with typical values falling in between (Miller, 2010). The closer the observed score aligns with the true score, the higher the reliability.

Similarly, G theory interprets reliability through the correlation between the observed score and the universe score. High correlation implies greater accuracy in predicting individual performance under generalized conditions. Unlike CTT's single error term, G theory differentiates various error sources, improving measurement accuracy (Shavelson & Webb, 1991; Brennan, 2001). The general form of a G coefficient is:

$$\mathbf{Coef\_G} = \frac{\hat{S}^2\mathbf{D}}{\hat{S}^2\mathbf{D} + \hat{S}^2\mathbf{G}}$$

Where $\hat{S}^2\mathbf{D}$ points to the estimated differentiation variance and $\hat{S}^2\mathbf{G}$ denotes estimated differentiation variance and the estimated generalization variance (Cardinet

et al., 2010). Differentiation variance includes components that define the object of measurement (e.g., persons), while generalization variance includes facets like raters, items, and occasions, adjusted by their sample sizes. The value of Coef_G depends on the nature of the decision being made (relative vs. absolute) and the complexity of the design.

Ultimately, G coefficients offer a refined, dependable metric for evaluating the precision and reliability of measurement procedures across multiple conditions (Shavelson & Webb, 1991; Cardinet et al., 2010).

In the application of Generalizability Theory (G theory), researchers such as Shavelson and Webb (1991), De Gruijter and Van der Kamp (2008), Meyer (2010), and Cardinet et al. (2010) identify three core reliability coefficients: the *relative G coefficient*, the *absolute G coefficient*, and the *criterion-referenced G coefficient* (Brennan & Kane, 1977). This study, like much of the literature, focuses on the first two.

The **relative G coefficient (Ep^2)** is appropriate for norm-referenced or relative decisions, such as ranking individuals. Developed initially by Cronbach et al. (1963) and further explained by Meyer (2010), it is defined as:

$$\mathbf{Ep}^2 = \frac{\boldsymbol{\sigma}^2(\mathbf{p})}{\boldsymbol{\sigma}^2(\mathbf{p}) + \boldsymbol{\sigma}^2(\boldsymbol{\delta})}$$

Note that $\mathbf{Ep}^2$ is the relative G coefficient denoting the mathematical expectation read as expected rho squared, $\boldsymbol{\sigma}^2(\boldsymbol{p})$ universe score variance, and $\boldsymbol{\sigma}^2(\boldsymbol{\delta})$ is the relative error variance.

In contrast, the **absolute G coefficient (Φ)**, or the dependability coefficient (Brennan, 2001), is suitable for criterion-referenced or absolute decisions, reflecting the precision of locating individuals on an absolute scale (Cardinet et al., 2010). It is given by the following mathematical formula:

$$\Phi = \frac{\boldsymbol{\sigma}^2(\mathbf{p})}{\boldsymbol{\sigma}^2(\mathbf{p}) + \boldsymbol{\sigma}^2(\Delta)}$$

(Φ) denotes the absolute G coefficient or dependability coefficient (read as phi

coefficient).

$\boldsymbol{\sigma^2(p)}$ represents the universe score variance; and

$\boldsymbol{\sigma^2(\Delta)}$ designates the absolute error variance.

The main difference between Ep^2 and Φ lies in their respective error variances, Φ includes more error components (facets and their interactions), making it typically smaller than Ep^2 (Myer, 2010; Bain & Pini, 1996; Pini, 2010).

The **criterion-referenced G coefficient**, Φ(λ)\Phi(\lambda)Φ(λ), introduced by Brennan and Kane (1977), estimates the accuracy of placing examinees in relation to a cutoff score. Its formula is:

$$\Phi(\boldsymbol{\lambda}) = \frac{\boldsymbol{\sigma^2 p + (\mu - \lambda)^2}}{\boldsymbol{\sigma^2 p + (\mu - \lambda)^2 + \sigma^2 \Delta}}$$

Where $\Phi(\boldsymbol{\lambda})$ reflects the criterion G coefficient (pronounced as Phi lamda).

$\boldsymbol{\sigma^2 p}$ refers to the universe score variance.

$\boldsymbol{(\mu - \lambda)^2}$ denotes the square difference between total mean score and cut off score.

$\boldsymbol{\sigma^2 \Delta}$ is the absolute error variance.

As emphasized by Marcoulides and Kyriakides (2010), G coefficients range from 0 to 1, with ≥ 0.80 typically indicating dependable measurement. Despite sharing foundational principles, the three coefficients differ based on their intended decision type (relative, absolute, or criterion), the structure of their error variances, and whether facets are considered random or fixed (Cardinet et al., 2010).

Up until now, in the design evaluation stage, three G coefficients have been discussed along with the corresponding sources of variance, what is left now is an examination of estimation precision measures accomplished via computing the standard error of measurement in G studies.

The standard error of measurement (Estimation precision)

The standard error of measurement (SEM) reflects the precision of estimation in scientific studies, acknowledging that random error is nearly unavoidable, even in well-

controlled designs (Shavelson & Webb, 1991). Although errors cannot be entirely eliminated, they can be minimized and their impact reduced. Lee and Cantor (2007) highlight that measurement errors affect the reliability of a procedure due to factors such as stability over time, consistency across forms, consistency in raters' judgments, and internal consistency. While G coefficients indicate how reliably conclusions can be drawn from assessments, the SEM measures the consistency of individual scores, emphasizing the variation between a student's observed score and true score due to error variance.

Cardinet et al. (2010) stress that the SEM quantifies the uncertainty and error impacting measurement results from Generalizability (G) studies, both in terms of relative and absolute measurement. It provides valuable insight into the reliability of measurement instruments. G theory also distinguishes between two types of error variances: relative and absolute decisions. Marcoulides and Kyriakides (2010) associate relative decisions with individual differences (e.g., ranking students) and absolute decisions with performance levels. For instance, assessing reading ability across different occasions and raters would be modeled in a p×r×o design, where rater and occasion are sources of measurement error.

The error variances for relative and absolute decisions can be computed using the following formulas:

$$\boldsymbol{\sigma^2 Rel} = \frac{\boldsymbol{\sigma^2 pr}}{\boldsymbol{n'r}} + \frac{\boldsymbol{\sigma^2 po}}{\boldsymbol{n'o}} + \frac{\boldsymbol{\sigma^2 pro,e}}{\boldsymbol{n'r\ n'o}}$$

$$\boldsymbol{\sigma^2 Abs} = \frac{\boldsymbol{\sigma^2 r}}{\boldsymbol{n'r}} + \frac{\boldsymbol{\sigma^2 o}}{\boldsymbol{n'o}} + \frac{\boldsymbol{\sigma^2 pr}}{\boldsymbol{n'r}} + \frac{\boldsymbol{\sigma^2 po}}{\boldsymbol{n'o}} + \frac{\boldsymbol{\sigma^2 ro}}{\boldsymbol{n'r\ n'o}} + \frac{\boldsymbol{\sigma^2 pro,e}}{\boldsymbol{n'r\ n'o}}$$

These formulas capture three principal components for relative error: student-rater interaction, student-occasion interaction, and residual error. For absolute error, there are six components: rater variance, occasion variance, student-rater interaction, rater-occasion interaction, student-occasion interaction, and residual error. As absolute error incorporates more variance components, it typically results in higher error variance compared to relative error, which is why the G relative coefficient is usually higher than the absolute coefficient.

The standard error of measurement (SEM) is essentially the square root of the measurement error variance, as described by Cardinet et al. (2010), who refer to it as "the square root of the measurement error variance" (p.31). G theorists call it the standard error because it quantifies the mean error in measurements, providing a standard measure of error. According to Bertrand and Blais (2004), the SEM is a theoretical construct that represents the standard deviation of the theoretical distribution of all potential errors under specific measurement conditions.

In practical terms, the SEM is an inherent random fluctuation that adds an amount of error to the measuring instrument or object being studied (Cardinet et al., 2010). Psychometricians like Brennan (2001), Shavelson and Webb (1991), and Myer (2010) have proposed formulas to calculate the SEM: $\sqrt{(\sigma^2(\delta))}$ for relative error and $\sqrt{(\sigma^2(\Delta))}$ for absolute error, which can be used to determine the SEM.

Cardinet et al. (2010) note key principles when calculating the SEM in G studies, emphasizing the "differentiation facet", the object of measurement. The SEM applies to one object of measurement at a time (e.g., students) and does not concern individual score differences. Its calculation is based on ANOVA-derived variance components. The differentiation facet can be either random or fixed. In G studies, random facets dominate, as they don't hinder research or generalization. The differentiation facet reveals how error impacts the measurement when replicated under similar conditions, contributing to both the error's magnitude and the SEM's square root.

When the differentiation facet is fixed, the Whimbey correction is applied to variance components, making them additive. This results in two SEM values: one "corrected" and one "uncorrected." The corrected SEM accounts for the error variance impacting each object of study, while the uncorrected SEM displays higher variance and represents the expected error variance with uncorrected components. These two values offer separate, valuable information for interpretation and conclusion drawing. G theory's EduG program is used to calculate the corrected SEM when the facet is fixed. For random models, the differentiation facet can be estimated without applying Whimbey, as it produces similar results for error variance estimates.

3.5. 5. Optimization (D Studies)

Generalizability Theory (G Theory) extends beyond assessing measurement accuracy and dependability; it also facilitates the enhancement of measurement procedures by analyzing sources of error variance. This enhancement process is encapsulated in the Decision Study (D Study), the final stage in generalizability studies. Researchers utilize data from the Generalizability Study (G Study) to identify optimal conditions for measurement procedures, aiming to achieve higher reliability and validity (Bertrand & Blais, 2004; Shavelson & Webb, 1991; Cardinet et al., 2010).

The G Study provides theoretical insights into sources of variance and error, but these insights are rendered practical through the D Study. The D Study employs "what if?" analyses to explore potential improvements in measurement procedures by adjusting facet levels, such as the number of raters, tasks, or occasions (Cronbach et al., 1997; Cardinet et al., 2010).

Optimization strategies in D Studies include:

1. Adjusting the number of facet levels: Increasing or decreasing the number of facets like raters or tasks can reduce error variance and enhance reliability (Shavelson & Webb, 1991).

2. Eliminating atypical facet levels: Removing facets that contribute disproportionately to error variance can improve measurement precision (Bertrand & Blais, 2004).

3. Modifying facet numbers and nature: Changing the number or type of facets, such as nesting items within objectives, can minimize error variance (Cardinet et al., 2010).

4. Eliminating measurement bias: Differentiating between students from different classes exposed to the same teaching method can reduce bias and enhance reliability (Cardinet et al., 2010).

These strategies enable researchers to optimize measurement procedures, ensuring that assessments are both reliable and valid.

To conclude, this section illustrated how far researchers can apply G theory framework to set up the general methodology in research studies. Principles and

concepts in G theory showed how a psychometric study can embrace multiple sources of error variance in G analysis simultaneously. Put differently, how one facet can affect the other in a single or multifaceted observation, estimation or measurement deigns. G theory can also optimize the dependability of a given assessment under specific conditions and constrains set for relative or absolute interpretations or decision making.

3.6. G Theory Designs for Data Collection

To estimate the variance components in G studies, many data collection designs have been proposed by psychometricians. These designs can be traced back early to the works of Gleser, Cronbach and Rajaratnam (1965), and later have been made simpler and practically exemplified within G analyses by other leading figures (e.g. Brennan, 2001; Shavelson & Webb, 1991; Webb et al., 2006; De Gruitjer & Van der Kamp, 2008; Meyer, 2010; Cardinet et al., 2010; Marcoulides & Kyriakides, 2010).

Inherent in G theory data collection is the term design whereby researchers can provide an estimate of the various sources of measurement error existing within the universe of admissible observations. If the universe contains one facet it requires one-faceted universe and if it takes more facets it, thus, implements a universe with two or more facets of measurements, or put simply a universe with multi-faceted designs. Selection of appropriate designs is subject to the nature of facet sampling, whether fixed or random or both (finite random). The sampling procedure would come up with random designs and fixed or mixed designs. If facet interrelationship with other facets is crossed or nested this will result in crossed and nested designs (either fully or partially nested. According to the founders of G theory (Cronbach, Gleser, Nanda and Rajaratnam), there exist two major types of relationships between facets of measurements, either “crossed” with one another, or one is “nested” in the other (Cardinet et al., 2010, p.13). The crossing or nesting facets interrelationships will be provided a detail explanation along generalizability study designs set up forth along the upcoming sections.

3.6. 1. Random Designs

As its name suggests, a random design depends on random facets that might be sampled from the universe of admissible obserations, where the universe may include

one or more facets of measurement.

3.6.1.1. One-Facet Designs

In G theory, a *one facet-universe* hints at a single faceted-universe that has one single facet of measurement; one source of measurement error. In this case, informants or individuals are often considered object of measurement or differentiated facets, symbolized as **(p)** and facet conditions symbolized as **(i)** that often refer to test items. When the two facets interrelationship is crossed, the sign **(×)** is used, once it is nested the relation is indicated by the symbol **(:)**, the resulting design would be either crossed **p×i** (the notation is read as person by items design) or nested **p:i** (read as person nested with items design).

These designs are assumed to rely solely on random facets whether for items or for other facets like occasion, test forms and raters. Items, be it multiple choice questions, is defined as a facet of measurement since it helps researchers and decision makers to generalize from a group of test items to a wide range set of test items. Items universe refers to all possible/admissible items that can sampled for study purposes. The same conceptualization can hold true for occasion universe and test forms universe. Facet random sampling is a feature so characteristic of single-facet designs because, as Meyer (2010) states, a two-faceted design might embrace fixed facet besides random facets of measurement.

3.6.1.1.1. One-Facet Crossed Designs

Designs with crossed facets have been defined by Shavelson and Webb (1991) as, "a measurement in which all conditions of one facet (e.g., items) are observed with all conditions of another source of variation (persons) as a crossed measurement "(p.11). An observation design, and measurement design henceforth, is crossed when "every level of one of the facets is combined with every level of the other in a data set" (Cardinet et al., 2010, p. 13). In an achievement test, for instance, we can consider a design that involves item and rater facets. When all test items appear in the exam sheet for all students to answer, and if all raters have to scan all the students' answer sheet for scoring and grading, then the two sources of variance of raters and items or students and items are crossed. One more illustration, when a group of students are required to

individually respond to all sets of tasks included in the test, the two facets of students **(p)** and tasks **(t)** are crossed, and this crossing partitioning is symbolized with a multiplication sign. For example, **(p×t)** or simply **(pt)** using Brennan's (2001) notational conventions, this crossing relationship would be read as person crossed with tasks. In other measurement situations, conversely, a number of facets, which are variance components or sources of error, can involve more than two facets, be it occasion **(o)** added to the previous design. By occasion is meant, students' performance over time; before and after training, in the morning and afternoon, different days of the week, after learning intervals, etc. A measurement design resulting is, therefore, three faceted where students tasks and occasions are crossed and symbolized as **(p×t×o)** or simply **(pto)**.

A one-facet person-crossed-with-items **(p×i**) design, the simplest design type, is seen as an analysis of variance ANOVA, a design consisting of a single random or fixed factor, in which test items is the only facet of measurement procedure and, of course, presenting one source of error or variability. G analyses designs considers three variance components of person variance ($\boldsymbol{\sigma^2 p}$), items variance ($\boldsymbol{\sigma^2 i}$), and person-by-items interaction variance as well as unmeasured random errors ($\boldsymbol{\sigma^2 pi, e}$).

It is important to pay some attention to occasion variability in G studies and designs. Variability due to test conditions when taken as an object of measurement causes a possible source of error in a measurement procedure and generalization since "generalization from sample to universe is hazardous" (Shavelson & Web, 1991, p.3). That is to say, test occasions vary in their circumstances for all students or even for some students; students having low performance on one sample of occasion may score high on another sample of occasions with the same or other test items. Test item facet and occasion facet, along other sources of variability, are two major challenging errors of measurement and generalization.

Shavelson and Web (1991) estimated four sources of error variance for a one-facet design. These are listed as under:

1-systematic differences among students achievement due to differences in knowledge and skills;

2-differences in the difficulty of the test items;

3-educational and experimental prior knowledge students bring to the test; and

4-randomness and unidentified events, for example different students administered the same test in different occasions and students' shift of attention respectively. (p.3)

In this kind of universe, the authors highlighted some challenging obsticles that are often related to the process of generalization. They represent both estimated systematic and unestimated systematic errors in this particular study, but this one facet design embodies part of the two and three-facetd designs set forth in our research. Our main concern is to address task, rater and theme main effects on score variability. For instance, variability due to task variety and complexity is a measured systematic variatoin and vaiability due to individual differences including conditions of passing the test, such as, knowledge that learners bring to the test, faigue, psychological status, noise in class, demotivation to take the test etc, is known as the residual component that signifies unmeasured random errors. The residual main effect is to be estimated but not to be controlled as a source of variation in our study even it might have a major effect on score variance and consistency.

In sum, one facet-crossed designs, as it seems, are very simple. They indicate that every individual from the very sample will respond to every single test items that are treated as random. That is, items are selected randomly from the universe of admissible observations among the possible other test items that all participants would answer. Subsequently, participants serve as object of measurement whereas items is a source of variance or variability.

As far as the nesting fatcet interrelationships within one facet universe, the following one facet netsted design is dedicated a detailed descritption being a foundation stone for planning our two G study designs.

3.6.1.1.2. One-Facet Nested Designs

A measurement model in G studies set for data collection can be nested. Nesting interrelationship between facets has been described by Shavelson and Webb (1991, p.46) in the ANOVA as "one factor (call it A) is nested within another (call it B) if (a)

multiple levels of A are associated with each level of B, and (b) different levels of A are associated with each level of B". This excerpt implies that the same sample of students, for instance, are required to answer different multiple choice questions, for example. In this case, test takers are not administered the same test items. The resulting design is **(i:p)** or **i(p)** read as items nested within students. This measurement model allows the researcher to estimate one single source of variability (Meyer, 2010), that of items in this specific measurement design.

As opposed to crossed designs, nested designs have received criticism for their limited utility in behavioral measurements. They have been proven to be ineffectual in drawing and generalizing conclusions as it decomposes the universe score into variance components far less than those of crossed designs. The variance components estimated in one-facet nested design are restricted only to two sources of variability: ($\boldsymbol{\sigma^2 p}$) person variance component and item variance interacted with person variance and variance due to unmeasured random error ($\boldsymbol{\sigma^2 i, pi, e}$ **)** component.

One more drawback adressed to the nested design lies in being less informative than the crossed design (De Gruijter & Van der Kamp, 2008). In fact, raters, the facet of measurement, in crossed designs allows the researcher to compare the results. Raters can be compared in terms of their severity or leniency because they rate all persons and all test items being administered to every students. However, in the context of nested designs, it is not possible to compare raters' ratings as they scored different items and different members within the same sampled group. Besides, the item effect or rater effect of the nested design is not differentiated within relative or absolute error variances.

The number of facets sampled from the univesre of admissible observations determines the type of G study designs, whether one faceted or two faceted, and thus for the type of interrelationship existing between facets, crossed or nested, in what follows multi-facet designs with two or more facets of measurement within crossed and/or nested status will be exclusively discussed .

3.6.1.2. Multi-Facet Designs

Behavioral measurement is more a complex process. Measurement in educational

sciences extends to encompass two or more facets (Marcoulides & Kyriakides, 2010; Shavelson & Webb, 1991). The stimulus for G theory implies that the major principle is not so much put on estimating reliability or, bettering generalizability, as on estimating the different sources of variability because information on the relative size of each estimated variance component explains the impact of the estimated variances on measurement error (De Gruijter & Van der Kamp, 2005). This shows that the sources of variability are put besides other interests at the very heart of G theory and G studies upon which D studies can be based. This rather justifies why designs with more facets are very effective in the process of data gathering to find out how well a measurement instrument is accurate and dependable.

3.6.1.2. 1. Two-facet Crossed Design (with Two Facet Universes)

A two-facet universe is commonly defined by two facets (e.g., test items and raters), and multi-faceted universe encompasses more than two facets. However, many sources of inconsistency emerge especially when several facets are sampled in a given measurement situation. Shavelson and Webb (1991) again highlighted other sources of inconsistency in behavioral measurements with two facets. These involve:

- Individual differences which is the universe score variability;

- Inconsistencies among raters due two variation in scoring;

- Occasion: having a test the day before spring holiday may affect students' achievement and their observed score would vary if passed their test at the beginning of the semester; here attention given to tasks is different.

G theory also studies the interaction between sources of error variance. This variation might induce researchers to examine the following issues:

a) person-by-rater interaction as object of measurement because "only some people and some raters in combination produce a unique result" (Shavelson & Webb, p.10). In some test settings a desperate need to consult dictionary in one occasion and not in the other is an issue that might be raised witin the context of investigating occasion variability effect on test score consistency;

b) Rater-by-occasion interaction: a source of variability in which a given rater can be

permissive for all students in one occasion and not in the other; and

b) The residual that contains the sequence of person, raters, and occasion (the person-by-rater- by- occasion interaction).

Along with this, this research considers students as object of measurement upon which it is believed valuable to draw accurate or 'unique' results. It also considers student rater interaction effects and the residual that embraces facets of students, raters, tasks and themes and their resulting interaction effects.

In G studies, there exist a number of facet universes types, and so for designs.

All encompancing two-facet crossed designs, two facet nested designs, and two facet mixed designs. The two-facet crossed design is the most frequent design applied in psychological and educational measurements. It is the most effective in the process of identification of sources of errors and their effects in G studies and in particular measurement situations. If students are assessed in their ability to use a sample of words correctly in their speech production to convey particular meanings, of course, with particular respect to grammatical boundaries and correct pronunciation, stress and intonation, etc, their speech production relevant to prompt tasks is rated individually by different judges using a given rating scheme, the appropriate design that goes for this case is composed of three fully-crossed facets, the notation is (**p×t×r**) for the person-by-task-by-rater interaction where students **(p)** are crossed with tasks **(t)** and with raters **(r)**.

The universe of admissible observations in this particular measurement situation involves two basic facets of tasks and raters and students are considered object of measurement. Since every rater should weigh every task performance, tasks and raters are crossed within the universe of admissible observations. However, students do not represent a component within the universe of admissible observations but a component for the population (Brennan, 2000).

The total observed score variance obtained from the speech production prompts can be decomposed into seven independent estimated parts called variance components. These components are supposed to be random effects variance components when the

population and facets in the universe of admissible observations are indefinitely large (Brennan, 2010). The seven variance components (or parameters) include both estimated variances: The estimated variances attributable to persons ($\boldsymbol{\sigma^2 p}$), rater ($\boldsymbol{\sigma^2 r}$), tasks ($\boldsymbol{\sigma^2 t}$) and interaction variance components: person-by- task interaction variance ($\boldsymbol{\sigma^2 pt}$), person-by-rater interaction ($\boldsymbol{\sigma^2 pr}$), task-by-rater interaction ($\boldsymbol{\sigma^2 tr}$), and residual variance ($\boldsymbol{\sigma^2 ptr, e}$). Estimating variance components for the two-facet crossed design in G studies can be conducted by quantifying the expected mean squares.

3.6.1.2.2. Two-Facet Nested Design

De Gruitjer and Van der Kamp (2008) and Shavelson and Webb (1991) distinguish four types of nested designs with two facets: **i ×(j:p**)- **j: (i×p) - (i×j): p - j:i:p**. The first **i×(j:p)** design is formally parallel to the **p×(i:j)** design, where the persons and one of the two facets have swapped places. In this particular design, individual responses are judged by a set of raters, where the set of raters differs from one person to the other. This partially nested design formula will be used in this particular research to denote the **p×(t:h:)** design that will be applied to investigate the relative impact of tasks and themes on score validity and consistency (see chapter five for a full description).

In the second nested design **j:(i×p)** formula, the **o:(i×p)** design where **'o'** denotes occasions. A clear case of the **o:(i ×p)** design is a design where every examinee responds to similar tasks, but occasions differ for persons as well as for tasks.

The third nested **(i×j):p** design is exemplified in the **(o×j):p** design in which a group of judges scores the work of a person, and each person is rated at different occasions and by a different group of judges. For instance, person 1 is judged at occasions 1, 2 and 3, by judges 1 and 2, person 2 is judged at occasions 4, 5 and 6, by judges 3 and 4, and so on and so forth.

The final, is a fully nested **i:j:p** design compared to the previous partially nested designs. In this measurement situation each person's performance is judged by a different group of judges, and each judge uses another set of tasks **i**. In this particular design only three variance components of person variance, rater variance and item

variance can be estimated due to the confounding of many effects. One advantage for this nested design lies in its efficiency to estimate condition effects (De Gruitjer & Van der Kamp, 2008) where facets interact. Rater 1 and rater 2, for instance, may overlap in one common group. They rate one group all together besides other diverse groups. Only in a single group that the rating effects can be compared in terms of severity and leniency and the effects they bring to individual scores.

As mentioned earlier, measurement in human and social sciences is complicated. G theory extends to universes with three or more facets when fitting the model to the measurement situation. Measurement variables entail many facets like items, occasion, and raters to intrude in the measurement process. Test takers' performance may vary across tasks, occasions, and via raters scoring. These facets being object of investigation in G theory can be used as measures for generalization. Speaking of facets leads researchers to tackle universes of admissible observations that also contain more than two facets but this procedure might lead to increasing generalization errors, more precisely said in Shavelson's and Webb (1991) words, "the broader the universe of admissible observations, the greater the possibility of making an error in generalizing from sample to universe" (p.10). Besides multi-faceted designs with more than two facets implemented in G studies and analyses, mixed designs contain random, fixed or mixed facet depending on the sampling status, the main theme that is considered under.

3.6. 2. Mixed Designs

As stated earlier in the estimation design section, developing appropriate measurement models do not only adhere to the number of facets specified in the universe of admissible observations, or to the nature of facets interrelationships (crossed or nested), but also involve facet sampling status, random, fixed, or finite random. This what defines the universe of generalization and the G theory model; that is, a G study design might contain purely random facets as they might involve mixed facets, random and fixed. These patterns are called mixed designs. Facet sampling status might include fixing a facet, and a fixed facet itself might influence the universe identification and increase score variance, and this would increase reliability in mixed designs in particular compared to random designs even imply equal facets and facet interrelationships

(Meyer, 2010). G theory implements a sampling theory to the improvement of measurement procedures (Cardinet et al., 2010, 18). Accordingly, fixing a facet is a measure that can be applied in D studies to improve the reliability of measurement procedures as it reduces fixed facet variance and increases G coefficients (Bain & Pini, 1996).

The estimation design stage discussed so far explains level sampling procedures used to determine the generalization universe and facet status. A mixed design in G studies should involve at least one single random facet (Shavelson & Webb, 1991) because G theory is a random sampling theory. As such it is not possible to model designs with purely fixed facets in this particular study as this never existed in G theory paradigm because its main concern was initially developed to investigate random effects. According to Cardinet et al. (2010), any measurement procedure can simultaneously carry out both fixed and random facets, and random sampling can also involve a finite universe. In random designs described so far, particularly one-facet universe designs, facets are always treated as random. Two facet universes, however, might involve one random and one fixed facets. Facet sampling status plays an influential role in the algorithms applied to estimate variance components and in the calculation of measurement error and G coefficients (Shavelson & Webb, 1991). Flexibility as a treat made diversified possibilities available for facet sampling status made the process of defining estimation design somehow critical and challenging for researchers.

In consequence, in mixed designs the variance is not computed in similar formulas either for fixed or random facets (Cardinet et al., 2010). For random facets, computing the mean square is based on dividing the sum of squares by the number of degrees of freedom, which is commonly equated to the number of squared values minus 1. For a fixed facet, on the other hand, the sum of squares is calculated divided by N, which refers to the number of levels in the facet population. Mixed designs for both random and fixed facets do not adhere to equal rationale (computations) as they produce distinct values; the former produces unknown value, whereas the other reports an observed result. So, it is not possible to combine the two values or variance estimates, because

they apply different computations and produce different variance estimates (Shavelson & Webb, 1991; Cardinet et al., 2010).

Subsequently, Whimbey, Vaughan, and Tatsuoka (1967), proposed the correction procedure to ensure comparability of the two computations of variance estimates. This correction implies the multiplication of the variance obtained by ANOVA by (N - 1)/N, a coefficient that relies on N, the population size. They confirm that in a fixed-effects variable the levels sampled include the total population of levels, and not a sample of size from a large amount of levels hence they suggest N-1 as the appropriate divisor instead of N of the sum of squares. In the same context, Cardinet et al. (2010) further comment that the correction procedure has no impact when sizes of population are infinite, and the contrary for small population sizes that can produce a difference. In the extreme case sampling, when facet levels equals (J = 2), the variance estimate that comes from a random sample is twice as large as when it results from a fixed facet. The *EduG* computer software provides an estimation for variance components for both random and fixed variance components, and so for the corrected variance components due to fixed facet levels effects.

To estimate the variance components in mixed models, Shavelson and Webb (1991) propose two methods. The first method implies averaging over conditions of the fixed facet in three steps. First, running an analysis of variance by treating all sources of variance as random including even those considered fixed. Second, identifying the random portion of the mixed design and the related variance components to be calculated; and finally, calculating the variance components for the random portion of the mixed design identified in Step 2, and each variance component of this source is added $(1/n)$ [6]. The second way is to analyze each condition of the fixed facet separately.

The resulting G coefficients values, when fixing a facet, are high. Fixing a facet is likely to increase score variance and reduce measurement error leading to an increased estimated reliability. Fixing a facet might alter universe score and limits the universe of generalization. It follows that the estimated reliability in mixed designs have different interpretations compared to random designs, where generalization in fixed designs becomes blurred (Meyer, 210). The decision maker might wish to shift an interest in

applying the principle of symmetry, where s/he might fix the facet of differentiation (object of measurement) as a special case (Cardinet et al., 2010). Hence, an alternative coefficient known $\boldsymbol{\omega^2}$ (read the expected omega squared) would substitute the $\boldsymbol{Ep}^2$. The major distinction between random and fixed facets lies in the idea that variance in objects of measurement in fixed facets is subject to direct observation, and the case is not true for random designs, as it should be estimated in infinite random facet.

Fixing a facet, and hence its effect, of measurement has a potential to increase the value for universe score variance and is likely to reduce the relative error variance and absolute error variance within a measurement situation. To illustrate, the analysis of data from a two crossed-facet design sometimes with random facets and sometimes with fixed facet (fixing occasion as an example) would lead to changes to occur at the level

Design	Variance	Formula
P × i × o **o, i random facets**	Universe score	$\boldsymbol{\sigma^2(p) = \sigma^2(p)}$
	Relative error	$\boldsymbol{\sigma^2(\delta) = \frac{\sigma^2(pi)}{n'_i} + \frac{\sigma^2(po)}{n'_o} + \frac{\sigma^2(pio)}{n'_i n'_o}}$
	Absolute error	$\boldsymbol{\sigma^2(\Delta) = \frac{\sigma^2(i)}{n'_i} + \frac{\sigma^2(o)}{n'_o} + \frac{\sigma^2(pi)}{n'_i} + \frac{\sigma^2(po)}{n'_o} + \frac{\sigma^2(io)}{n'_i n'_o} + \frac{\sigma^2(pio)}{n'_i n'_o}}$
P × i × o **i: random facet** **o: fixed facet**	Universe score	$\boldsymbol{\sigma^2(p) = \sigma^2(p) + \frac{\sigma^2(po)}{n'_o}}$
	Relative error	$\boldsymbol{\sigma^2(\delta) = \frac{\sigma^2(pi)}{n'_i} + \frac{\sigma^2(pio)}{n'_i n'_o}}$
	Absolute error	$\boldsymbol{\sigma^2(\Delta) = \frac{\sigma^2(i)}{n'_i} + \frac{\sigma^2(pi)}{n'_i} + \frac{\sigma^2(io)}{n'_i n'_o} + \frac{\sigma^2(pio)}{n'_i n'_o}}$

Table 3.2: Contrasting Universe Score, Relative Error and Absolute Error Variances within Random and Mixed Designs in p×i×o (Meyer, 2010)

of the universe score variance and error variance mathematical formula as well, simply because they serve basic components resulting from calculations corresponding to mixed designs. These modifications are displayed in Table 3.2 above.

Form the above table, it is noticed that the person-by-occasion interaction component appears within the universe score variance in the mixed design; however, it is totally eliminated from the relative error and absolute error variances. Additionally, occasion variance is deleted from the absolute error variance. In consequence, reliability in fixed models is expected to be higher than that obtained from random designs.

Similarly, in the **p×(i:o)** two-facet mixed design with **"o"** considered fixed, both mixed and random designs lead to equal changes at the level of the universe score variance, the relative error variance and the absolute error variance. These are statistically best explained along the table under:

Design	Variance	Formula
p ×(i : o) o, i random facets	Universe score	$\boldsymbol{\sigma^2(p) = \sigma^2(p)}$
	Relative error	$\boldsymbol{\sigma^2(\delta) = \frac{\sigma^2(po)}{n'_o} + \frac{\sigma^2(pio)}{n'_i n'_o}}$
	Absolute error	$\boldsymbol{\sigma^2(\Delta) = \frac{\sigma^2(i:o)}{n'_i} + \frac{\sigma^2(o)}{n'_o} + \frac{\sigma^2(po)}{n'_o} + \frac{\sigma^2(pio)}{n'_i n'_o}}$
p ×(i : o) i: random facet o: fixed facet	Universe score	$\boldsymbol{\sigma^2(p) = \sigma^2(p) + \frac{\sigma^2(po)}{n'_o}}$
	Relative error	$\boldsymbol{\sigma^2(\delta) = \frac{\sigma^2(pi:o)}{n'_i n'_o}}$
	Absolute error	$\boldsymbol{\sigma^2(\Delta) = \frac{\sigma^2(i:o)}{n'_i n'_o} + \frac{\sigma^2(pi:o)}{n'_i n'_o}}$

Table 3.3: Comparison Between Universe Score, Relative Error and Absolute Error Variances within Random and Mixed Designs in p×(i:o) (Meyer, 2010)

Table 3.3 can be interpreted according to Meyer's (2010) conceptualizations. He asserts that facet fixing might not increase the universe score variance and reduce error variance since it is associated with the design where a facet is fixed in. If items facet is fixed, the universe score and error variance formula changes in **(p×i×o)** especially when compared to **p×(i:o)** design. That is, the universe score and error variances are equal in the random design **p×(i:o)** and in the mixed design with fixed items. Theoretically speaking, if occasion facet is infinite random and items are nested with occasion, items thus are, obligatorily, random. That is why **p×(i:o)** design with items as fixed

decomposes into the variance components similar to those involved in designs with randomized items.

All things considered, flexibility of G theory has been proven <with diversified data collection designs with random or fixed models along with designs with one, two or extended universes. Facet status and sampling conditions also represent the extent to which this random sampling theory is extensively adaptable to researchers' changing needs and objectives.

3.8. Generalizability Theory in Language Testing

Although extensive research conducted on vocabulary teaching, vocabulary learning strategies, lexical coverage in written and spoken texts (lexical sophistication, lexical diversity, lexical density and lexical richness), academic vocabulary and data bases vocabulary assessments, the impact of vocabulary knowledge on the four language skills and overall language proficiency (see chapter 1 for a full description), little attention seems to be paid to investigate consistencies and inconsistencies of vocabulary test scores, obtained either from size or depth tests, using G theory principles and procedures. That is, issues of assessing the assessments is irrelevant to an examination of vocabulary measures. The application of G theory is critical to studying the dependability of test scores and the usefulness and feasibility of inferences and interpretations drawn, upon which norm-referenced or criterion-referenced decisions are being made for placements tests or for other purposes.

In the domain of language testing, a number of studies have been published applying G theory, particularly in the context of performance testing from different perspectives, notably, in listening, speaking, reading and writing. Much of work within the framework of G theory was driven to studying task and rater effects as well as person task effects in L2 performance. Because of the outnumbered studies carried out in the field, we opted for discussing only the studies that come up next.

Barkaoui (2007) used G theory to investigate the relative effect of holistic rating scales and rating scales with multiple traits on the scores of L2 writing performance obtained from four novice, not trained scorers. The G studies adopting the holistic scale, revealed that the largest variance component contributing to score variance was the

person-by-task-by-rater interactions by (61.59%), pursued by students' main effects by (26.23%), raters by (5.73%), person-by-task interactions by (5.01%), and task-by-rater interactions by (1.43%), with negligible values for tasks (0.00%). and person-by-rater interactions (0.00%), thus the latter facets of measurement show no impact on the writng scores obtained. G studies conducted on the impact of multiple traits scales indicated that most of the variance was attributable to raters by (58.18%), pursued by person-by-task-by-rater interaction variance by (37.58%), and students by (4.24%). The findings also displayed that there was no impact on score variability due to tasks and students-by-task, student-by-rater, and task-by-rater interactions with (0.00%) per each facet and its interaction effects. The findings further concluded that the ratings resulted from the holistic rubrics are dependable rather than multiple trait scoring.

Lee and Kantor (2007), in their work on evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory, investigated the relative effect of various rating designs and the number of tasks and raters on the reliability of 488 EFL students' writing scores based on TOEFL integrated, listening-based and reading-based, and independent tasks. Participants were grouped into three groups (n_p = 162, n_p = 164, n_p = 162 for each subgroup respectively); each subgroup responded to a different category of tasks containing six writing tasks out of eight tasks. Univariate and multivariate analyses with different crossed and nested models were implemented for data analysis. The results revealed, among other things, that dependability of the writing scores increased when the number of tasks increased than increasing the number of rater. Besides, the largest source of variance was attributed to the person-by-task interaction (by 20.3%) and the rater effect was relatively low compared to the total of score variance. The results recommend implementing integrated writing tasks on the TOEFL tests.

Gebril (2009) investigated the score generalizability of academic writing tasks. A number of 115 Egyptian university students were required to write essays on two independent tasks and two reading-to-write tasks. Three trained raters scored the essays holistically. A fully-crossed univariate design (p×t×r), and GENOVA were implemented as procedures for data analysis. Major findings illustrated that reading-to-

write tasks generated reliable scores and the same does hold true for independent tasks, and that generalizability decreased when only one writing task was prompted due to the student-by-task variance component.

Lane and Sabers (1989) applied G theory to study the reliability of writing scores gained from fifteen essays that were sampled randomly from three and eight grades students who responded to a single writing prompt. These essays were rated in terms of ideas, development and organization, sentence structure, and mechanics by eight judges. The study findings revealed that the generalizability coefficient ranged from 0.68 to 0.90 with raters number increased from one to four. Furthermore, the main source of variance attributed to the total error variance was the student main effect and the interaction between persons and tasks compared to the rater effect which was relatively small. Reliability increased when one rater to two raters were deployed, thus the writers caution the generalizability of scores because only a single prompt had been assigned for completion of the writing activity in the research.

In an exploratory study, Polat and Turhan (2021) compared nested and crossed generalizability models to investigate the consistency of scores gained from an assessment of 116 intermediate level FL learners studying at a Turkish state university's language school FL speaking skill. Precisely, the study aimed to compare G and Phi coefficients obtained from the scores of a full factorial, fully crossed, design contrasted to a nested design where rating scales were nested in four raters. Findings pointed out that the G and Phi values obtained with the full factorial grading design were higher (G: 0.85, Phi: 0.78) than the nested design (G: 0.79, Phi: 0.72). Besides, the major source of variance was attributable to the student's main effect of the full factorial model while the variance due to the residual components was lower. This means that the full factorial design could produce more reliable results in language speaking assessment especially if the number of raters is obtainable.

Burton (2008) investigated the impact of four facets of passage, day of test administration, rater and rating and occasion, to refine a rating procedure used to assess 28 intermediate elementary school fourth grade students' ability to orally retell what they had read from two expository passages using a partially nested design. G study

research findings arrived at six largest sources of variances all-encompassing students (33.9 %), student-by-day interaction (10.2%), the interaction of passage with rater nested within student and day (23.3%), the student-by-day-by-rating occasion interaction (3.4%), the passage-by-raters interaction (23.3%) and the residual (11.2%). D studies concluded that adding another reading day, as another occasion, would maximize reliability of the measure, rather than engage students in reading more passages, or induct more raters or more rating occasions. To increase score reliability students need to read two passages on at least two separate days and then have their performance rated by two judges across the various passages, testing days and raters.

Kumazawa (2009) administered a paper-and-pencil multiple-choice diagnostic criterion-referenced achievement vocabulary test for a group of 131 learners belonging to a general English course specialized in literature, law or economics at a high-ranking university in Kanto. This research examined the variance components that assessed the extent to which learners had receptively mastered the words pre-taught in the textbook after having four reading and listening sessions. Twenty five target words were randomly sampled from each of the five chapters that were, in turn, selected from 10 textbook chapters. From a total of 25 words, 5 items were nested within each section. Target words were underlined and embedded within sentences similar to the one participants had been previously exposed to in the textbooks. Learners were instructed to select appropriate meanings to the prompt word from a set of four choices, among which three were distractors that were selected from high-frequency or academic word lists. The results showed large variance due to item effects and no variability due to persons. Added to that, a lower level of items and sections in the MCT, would yield in unsatisfactory dependability (0.30), proposing a revision for test items. Increasing the number of items contributed to higher dependability and increased variability in students' scores; 40 items are needed to obtain (0.41) dependability index, and this affects cost effectiveness; or time consuming. It is suggested to teachers to consider time allotment.

As can be noticed, few previous studies have researched the dependability of vocabulary measurement. To address this research gap, G theory will be applied to

examine the dependability of vocabulary assessment scores, and the relative impact of sources of variance on score consistency and accuracy. As a matter of fact, there are few studies available that report on the reliability of vocabulary assessment scores assigned by raters to open-ended communicative responses within the task-based framework.

3.9. Contribution of Generalizability Theory to Research Generalization

No matter how research is carried out, its findings are subject to some or many limitations. To shed light on these research limitations and to reduce them, and or to identify sources led to such limitations, G theory came into the research scene. Social sciences researchers and evaluation practitioners are helped with its statistical instruments proving the consistency of behavioral measurements.

Proponents of G theory like Shavelson and Web (1991) stress the significance of assessing the major sources of error variance in order to decrease unsystematic variation. Perhaps the best words to support this section are said by Shavelson *et al*. (1989, as cited in Marcoulides & Kyriakides, 2010): "instead of asking 'how accurately observed scores reflect corresponding true scores', Generalizability Theory asks 'how accurately observed scores permit us to generalize about persons' behavior in a defined universe" (p.222). In this sense, G theory sounds effective to address issues of dependability and generalization of test results.

G theory is, by no means, the most efficient method for estimating multiple sources of error in measurements in a separate manner, in a single analysis. It allows for the estimation of the amount of measurement error. It helps to know the extent to which results obtained from G study and D study are accurate.

Gibril (2009) sums up the advntages of G theory and describes them as follows:

- G theory provides information about different sources of error as it is able to partition the error term into different part;

- G theory provides a unified approach to measurement error regardless of the facets involved in the assessment process;

- G theory provides a unified approach for assessing the reliability of measurement

designed for making either relative (norm-referenced) or absolute (criterion-referenced) decisions; and

- G theory makes no assumptions about the overlap of different sources of error, but simultaneously estimates these different sources.

Thus far, G theory has justifiably proved to be effective is assessing assessment procedures including tests, questionnaires, portfolios, observation grids and many other behavioral measurements. It targets quality assessment and addresses psychometric issues such as estimating the magnitude of error variance and its impact on score reliability and validity, and assessment precision henceforth.

Conclusion

In this chapter, G theory has been discussed in connection with the shortcomings of CTT. Then, a historical account of G theory with fundamental concepts that cut across CTT and G theory have been explored in detail, and the major principles grounding on G theory have been explained with regard to an investigation of consistencies and inconsistencies of test results and variance components attributable to measurement error.

The chapter has also overviewed crucial stages in the application of G theory. When conducting a G study, in observation design stage in particular, researchers need to specify as much as possible facets, their observed levels and interrelationships for purposes of generalization and error minimizing. In estimation design phase, researchers have to specify the sampling status of facets whether fixed or random. In measurement design, the starting point for generalizability analysis, the researcher identifies the object of measurement as a study focus, and decides whether a measurement is to be relative criterion-referenced or absolute norm-referenced. In the design evaluation stage, calculating G coefficients and the SEM is important to estimate dependability of scores and measurement precision. In a D study design, a researcher seeks to estimate optimal conditions set for a measurement procedure to yield in successful generalizability.

This chapter has further explored data collection G study designs with different

facet level universes and sampling procedures set for random, mixed and nested designs, proposed to estimate variance components affecting reliability of measurements. Finally, the chapter has thrown light on the sources of errors that affect vocabulary measuring procedures and various studies investigating vocabulary assessment using G theory and the contribution of G theory to the research endeavor have also been reviewed.

The first part of this thesis is devoted to the review of literature displayed in three distinct but related chapters dealing with vocabulary assessment, assessing the assessment, and G theory respectively. The second part of this research work is, therefore, dedicated to the empirical investigation of the reliability and validity of a vocabulary performance assessment using G theory principles. It embraces three chapters: Chapter four elaborates on the methodology and study design and procedures. Chapter five reports the research results and analyses data collected from the test of vocabulary knowledge using the EduG software package. Lastly, Chapter six discusses and interprets the research findings upon which a number of implications are suggested, perused by limitations of the study.

CHAPTER FOUR

RESEARCH DESIGN AND PROCEDURES OF DATA COLLECTION AND ANALYSIS

Introduction

This practical chapter begins with the description of the research methodology applied in the current research, which is psychometric in principle. It then argues for the design and procedures to be used for data gathering and analysis. Afterwards, it describes the sample of the study. Then the procedures used for test design and development are exposed in five steps: purpose of the assessment; content specifications; test structure and development; reviewing the psychometric properties of the research tool to establish content and construct validity evidence, test administering, and finally quantifying examinees' lexical performance. This chapter is further devoted to describe the different stages of G theory applied to collect and analyze the data as illustrated in the G study designs and the utility of *EduG* computer program. Thus, fittingly to the research purpose and context, a descriptive approach involving quantitative methods is opted for the data collection and analysis and to answer the research questions. In order to trace evidence for the dependability of students' obtained lexical performance scores, we opted for G theory, a statistical procedure for quantitative data-based investigation, to which we devoted a thorough discussion in the introductory chapter of this thesis.

For psychometricians, G theory is the best method used to investigate the psychometric characteristics of test scores, namely validity and reliability, the core issue of this study. Put otherwise, it seeks to explore the relative contribution of sources of error and variability on the reliability of the measuring instrument; it is an attempt to assess each variance component that has a potential to feature out the measurement procedure and hence improve its design in the D studies conducted. The research project is carried out at ENSB where a sample of 113 first year EFL students sit for a vocabulary performance test, whose obtained test scores might reflect their real level of performance and, it is suggested that these ratings serve as a corpus for the present study.

As clearly indicated in Chapter 1 devoted to the conceptual framework, vocabulary knowledge seems to be part and parcel in EFL learning and teaching, it is practically and theoretically difficult to measure, it reveals to be of a highly abstract construct. This is mainly due to its critical nature that made it thought-provoking to define; being a multidimensional concept all-encompassing form, meaning, and use underlying receptive and productive sub-knowledge, this construct proved to be difficult to handle. Modelling its assessment, therefore, sounds tricky and challenging to over control especially that no convention has yet been made on its assessment formats. This is, in turn, what might cause variability, inaccuracy and inconsistency of scores obtained from its measurement procedures that might be affected by various sources of error when coming to assessing its assessment. Consequently, the study's main interest is to estimate the possible sources of error variance that might have a certain impact on vocabulary assessment, the research measurement situation. In this regard, a number of questions have been raised:

4.1. Research Questions

1- What is the relative effect of tasks and raters on the generalizability (for relative decisions) and dependability (for absolute decisions) of scores obtained from a vocabulary performance test?

2- What is the relative effect of tasks and themes on the generalizability and dependability of scores obtained from a vocabulary performance test?

3- What is the relative effect of tasks, raters and themes on the generalizability and dependability of scores obtained from a vocabulary performance test?

4- What is the effect of decreasing the number of tasks designed to assess vocabulary performance on the generalizability and dependability of test scores?

5- What is the effect of decreasing the number of raters on the generalizability and dependability of test scores?

6- What is the relative effect of tasks, raters, and themes on the construct validity of a vocabulary performance assessment?

These questions will be answered owing to the data gathered from the study statistical measures used for data collection designs and data analysis via conducting G study analyses and D studies. The steps set forward and the principles of G theory applied are to be described in subsequent sections of this chapter.

4.2. Research Method

In the third chapter of this thesis, we mentioned that the nature of this measurement study is psychometric in principle. In fact, when the research fundamental objective is to estimate the psychometric conditions of test scores, when measuring performance, a quantitative method sounds better indicated rather than a qualitative method. We indulged into gathering data about students' performance in an-depth productive vocabulary knowledge test (DPVKT) including a set of communicative tasks; these obtained statistical/numerical databases do exclusively derive from students' grades, but do not drive from observation of students' experiences or perceptions. Actually, information that feeds this research derives from students' performances displayed in response to the test instructions.

Noticeably, such information sounds quite consistent with what some researchers agree to be the core of quantitative data, which serve as evidence collected in a quantitative based approach. According to Aliaga and Gunderson (2002, cited in Muijs, 2010, p.1), quantitative research is "explaining phenomena by collecting numerical data that are analyzed using mathematically–based methods". Quantitative research manipulates pre-existing numerical data implementing computational techniques (Babbie, 2010). These authors, it is noticeable, highlighted statistical information both in terms of collection and analysis.

The abovementioned sources emphasized two features of quantitative data: one is associated with the data proper that are mostly numerical in nature and indirectly (nonverbal) obtained from respondents' responses to the various test tasks, and from raters quantifying or weighing those responses. The second feature is that computer softwares are used to analyze numerical data. In the actual case the EduG software package was applied. This suggest that the results are expected to be accurate and thus truthful to draw further generalizations.

Further to sounding more appropriate for exploring mathematical and numerical information, a quantitative-based approach is also favored for its utility and adaptability to both the purpose and context of investigation. G theory, being a statistical procedure used for investigating the dependability of measurements, is said to be an alternative method to CTT, which has proved to be convenient for examining the reliability and validity of alternate assessment scores (Shavelson & Webb, 1991). This statistical framework allows us to identify multiple sources of measurement error by means of a set of principles. It also permits us to identify relative and absolute generalizability coefficients in addition to searching and modelling thoroughly the facets of measurement that contribute to increasing or decreasing the generalizability and dependability of students' scores in a performance on the test items.

Given that the research aim is typically related to establishing the reliability and validity of the behavioral measurement for the present research, the American Psychological Association, and the National Council on Measurement in Education (1999) stressed a desperate need to rely on G theory in this respect. Our decision to opt for quantitative rather than an exclusively qualitative approach, was somehow supported by psychometricians' views on the suitability of applying G theory to examine measurement precision thanks to the various advantages it offers. These are listed by Cardinet et al. (2010) along the upcoming lines:

> Relative item difficulty, the mastery levels characterizing different degrees of competence, the measurement error associated with estimates of population attainment, the progress recorded between one stage and another within an educational program, the relative effectiveness of teaching methods, are all examples of G theory applications that focus on something other than the differentiation of individuals. To facilitate an extension to the theory, calculation algorithms had to be modified or even newly developed. (p.3)

The main concern of the present study is to estimate the consistency of students' scores across tasks, raters, and theme (vocabulary use in different contexts), and interaction of variance components (person-task variance, person-theme variance, person-rater variance, task-rater variance, theme-rater variance, person-task-rater variance, person-task-theme variance, and many other variances) through G studies analyses. It is important to mention, here, that analysis of the relative contribution of

each of these modeled sources of measurement error will occur simultaneously reflecting one merit of G theory over CTT. The study further attempts to draw inferences on the test scores across the applied assessment method and conditions (facets) and their interaction effects. Furthermore, the D study stage is expected to find out the reliability of scores that might be under variation in the present assessment design with eight communicative tasks and two raters. It provides approximations to increase the reliability index or consistency of scores either, for example, by increasing or decreasing the number of tasks, or by means of decreasing or increasing the number of raters.

With regard to the research design, the utility and appropriateness of quantitative-based data collection have been examined and its applicability to the research context and purpose have been arguably justified. The best approach which is logically relevant to yield mathematical information generated via computerized calculated algorithms implementing G theory, is therefore descriptive, where study participants' are measured once unlike experimental approaches where examinees sit for pre-tests and post-tests. Added to that, as most of psychometric studies are exploratory, this approach is meant to explore the sources of error variance that might affect validity and reliability of students' observed scores. Based on these discussions, we argued for the need to document our study with quantitative data by considering both administering a test to gather numerical information and using specific data collection designs grounded on G theory. In what follows, the data collection procedures set forward to meet these research objectives are presented.

4.3. Data Gathering Procedures and Research Tools

Descriptive in nature, the research aims to estimate the reliability and validity of test scores obtained from an assessment of students' vocabulary performance. Given the set of concepts and principles, which have emerged from the literature on G theory, we realized that to address the question about the psychometric properties of test scores, we could not only rely on G study designs and analyses but also conduct a set of D studies. Thus, conducting the G study starts with collecting quantitative data to aid us to first explore the possible sources of error variance and then elaborate on in the D

studies to obtain optimal G coefficients. In essence, quantitative data depicted in a form of measurements and scores are gathered by means of descriptive and observational testing procedure.

So, students' depth of productive vocabulary knowledge was believed essentially useful in providing the study with information on the theme, assessing the assessment. In order to answer the six research questions deriving information about validity and reliability issues, we were of the opinion that much information about such hint may appear in the test. Accordingly, eight complex communicative performance tasks have been designed corresponding to the program of third year secondary school level. This program, on which we based our test, contains a set of themes that were believed to be convenient to meet students' exit profile and ensure fairness; giving equal opportunity to students, having equal prior knowledge, to pass the test. The test being constructed underpins four mandatory topics: Ancient civilizations, Education in the world, Ethics in business, and Feelings and emotions. It seeks to assess first year EFL university students' vocabulary knowledge from the perspectives of word meaning, word formation and word use in written discourse contexts; it targets their lexical ability within task-based performance.

Before having developed and administered the eight communicative tasks to learners, it was necessary to ensure their validation. Consequently, a checklist of construct validity was designed in quest of validating the test tasks from the part of EFL teachers, and a questionnaire was addressed to learners to refine the structure and content of the tasks (see sections below for a full description).

Preceding the report on the pilot phases of the devices here is a brief description on the test from G theory standpoint. It is a paper and pencil test that samples eight tasks from the universe of admissible observations. The tasks are selected from an indefinitely large number of tasks that have a potential to be items assigned to assess students' vocabulary performance. The current measurement situation, which involves a set of measurement facets, namely tasks, raters, and themes, treats students' observed scores (performance) from different perspectives; how far sources of measurement error can affect examinees' performance. For illustration, as far as the facet of rater is concerned,

this assessment is an attempt to explore the relative impact of raters' leniency or severity on students' observed scores. As to task and theme facets, the study explores whether students' performances change over tasks and themes, once the task or theme changes. As such, the data will be observed from three different but related perspectives: from facets of raters, tasks and themes. It is important to note here that even raters and themes are sampled from the universe of admissible raters and themes. In other words, two raters have been selected from an indefinitely large number of secondary school teachers in Algeria that can possibly be participants in quantifying students' performances. The same does hold true for themes that are sampled from thousands of themes existing in EFL vocabulary teaching.

Before providing a full description of test structure and design employed in this research project, the sampling procedure is considered first.

4.4. Sampling

To collect relevant data, we addressed students' who, according to the Baccalaureate results they obtained and their admission to the English course in the Department of English at ENSB were believed to be useful to draw on accurate results and reveal high achievement in vocabulary knowledge. We then thought it appropriate to observe their behavior or ability in a performance test, depicting what those students can do with that knowledge of words that, presumably, grasped after a course of instruction; during third year secondary school education. What they actually reached or attained after this year learning vocabulary exploring textbook units, and after three months vacation succeeding their BAC graduation, is the major concern of the study.

The need to pay attention to vocabulary knowledge, and word use or fluency in particular, spurred from a belief that they might be of some use to our exploration of the type of lexical knowledge that smoothes students' entrance/joining the teacher training school course at ENSB. Identification of this sort of knowledge will determine whether students are ready to join this course in the Department of English and succeed to develop certain competences as pre-service teachers to function adequately, afterwards, in the classroom when inducted into a new middle/secondary school.

Interest in the new bacalaureate holders enrolled in the teacher training school avenue to success suggested the need to identify their abilities, prior knowledge, and their readiness to take a new course, and watch over the type of responses they displayed in the test. The behavior being investigated, vocabulary knowledge, is elicited via four themes, those mentioned above, originating from the four mandatory units of the textbook *"new prospects"*. Interest in administering various tasks including various topics suggested a desperate need to examine theme relative effect on score performance variance.

This knowledge being explored will reveal whether receptive vocabulary knowledge is fully or partially transferred into productive use ability. Being exposed to similar syllabus content, and being accepted to register in this new course have led to the assumption that the students might prove to be different, due to individual differences, in their knowledge of word meanings, word parts and word use; it might thus partially explain sources of variability attributed to measurement error. This means that variance due to vocabulary performance might account for measurement reliability.

All in all, 113 first year EFL students enrolled in ENSB were purposefully selected; all the four groups were selected to achieve research purposes. At the beginning, the sample was larger it consisted 143 students. Yet, many students either were absent, because of the strike they did not yet join their classes, or did not respond to all the eight tasks or even did return the exam sheet in blank, or wrote irrelevant responses among others that refused to participate as they were given a choice to have a sitting exam. Hence, only those examinees who responded to all the tasks are accepted to participate in the study, since it endeavors to investigate the test takers' vocabulary knowledge throughout the whole course of instruction; across the four mandatory units of the textbook *"New prospects"* devoted to teach third year in secondary education.

In addition to the students' sample, two secondary school experienced teachers participated in the study to shed more objectivity on the obtained results. Being acquainted with knowledge about the examinees and the four textbook units, these teachers were believed to be handy to provide accurate results and, hence, they were in charge of evaluating the examinees performance across the eight tasks. At first three

raters were expected to score the test, unfortunately, however, one of the raters did not give back the exam sheets, and thus was excluded from the study sample. Participated in the study, one rater taught for more than 15 years the other for more than six years.

Having different levels of professional experience, the ratings obtained might rise bias and affect assessment precision. This discrepancy in experience might impact their interpretation of the scoring criteria. To avoid rater bias, inter-rater experience and reliability considerations were taken seriously into account by:

- Training and calibration: we ensured that both raters were trained on the same scoring rubrics and practice scoring within the same tasks and have discussions afterwards;
- Desining clear rating rubrics with detailed and objective criteria; and
- Pilot testing: both raters scored the same set of tasks and we calculated inter-rater rliability.

These were the procedures posed to control and minimize the risk that might emerge from raters' judgements and drift.

It is also worthmentioning that a number of participants were also involved in the pilot studies, conducting in quest of test validation purposes: 1) seven teachers reviewed the first draft of the checklist of content validity containing the tasks eight prompts for further item writing; 2) 22 teachers expertized the performance tasks included in the final draft of the checklist of content validity to further review the psychometric properties and credentialing the test content and format; 3) 50 first-year students responded to the first draft of the test to elicit their performance, validating its content and trying out the scoring guide; 4) 33 first-year students answered a survey questionnaire to check their attitudes towards test item types complexity, instruction clarity, cognitive load, authenticity, …etc. to further refine the tasks structure when necessary by means of reviewing the psychometric properties; 5) and finally, a project group composed of five teachers developed the scoring guide in order to quantify observation.

Now that the sample have been delineated, the different procedures employed in designing performance tasks for the data collection will be examined next.

4.5. Designing Performance: Test Design and Development

In Chapter 1, we concluded with reference to the challenging obstacles that researchers might face when designing vocabulary depth tests that, there is no such flawless model of vocabulary assessment to follow in this study, as its assessment instruments are still under search for development especially that this test translates performance based-assessment principles. That is to say, the principles of reliability and validity (see Chapter 3) are also under research question as many other performance tests' psychometric properties are. Despite this fact, there is an attempt from the part of the researcher to construct a test, which aims at gathering information about first year EFL degree students' lexical performance within performance-based assessment framework.

During the process of test design and development, four significant resources have guided the development of a framework of vocabulary performance assessment followed in the present study. These are: Genesee's and Upshur (1996) book entitled *"Classroom-based Evaluation in Second Language Education*" that was informative of the major principles of language assessment, AERA, APA, and NCME (1999) assessment standards whereby test content was defined. Read's and Chapelle (2001) vocabulary validation framework that had also guided us to consider validity and reliability issues on vocabulary assessment and make decisions about test format, and above all Johnson's et al. (2009) system approach that sound very consistent with performance assessment. This approach enlightened us with decisive steps used to design and administer the assessment. It was largely adapted, when compared to the other resources, to serve the current research work evaluation plan. In essence, the current assessment system revolves around four basic steps: Delineating the assessment purpose; defining the test framework that stresses the domain of the assessment or the research construct; developing prompts and structuring tasks that will elicit the students' performance on vocabulary together with their expected responses; and finally developing a scoring guide for weighting performances. These steps are further explored in the following sections:

4.5.1. Purpose of the Assessment

The current performance assessment purpose is to measure students' ability to use previously acquired knowledge in solving novel problems or completing specific tasks, particularly, word use in written discourse context . This alternate assessment aims at identifying examinees' qualification for graduation from secondary school from a vocabulary knowledge perspective to further provide descriptive information about student's achievement in a vocabulary performance test to join a new course at university. This measurement is an attainment test that occurs after a course of instruction. As the test is summative in nature, it aims to certify mastery of the units' objectives, using target simulated words. It investigates whether examinees can demonstrate their ability to use words being internalized or recognized, i.e., testing the ability to transfer their passive knowledge to active or productive knowledge. Thus, students' development of lexical competence is elicited via writing short paragraphs about a number of topics situationalized in authentic settings using some target words that had been previously taught.

Now, one might wonder why the application of G theory is being considered in the research work. The reason for this authentic performance assessment is to illustrate how G theory is effective and can be applied to determine multiple sources of measurement error and assess the reliability of test scores from various assessment designs based on the facets of persons, tasks, raters and themes.

A preliminary step in designing the assessment was to establish the assessment purpose, the next step is then devoted to develop test specifications that identify the research construct, knowledge, skills and abilities intended for the current measurement.

4.5.2. Development of Test specifications

According to AERA, APA, and NCME (1999), the development of test specifications requires a detailed description for a given test. Enlightened by this insightful source, the present vocabulary assessment framework sets forth the construct of vocabulary knowledge along three aspects: test framework, test specifications and a specification table. Each aspect of these is considered next.

4.5.2.1. Test Framework

Test framework, or what Genesee and Upsher (1996) referred to as specifications for test validators represents the theory of learning in this study, which is either clearly stated by curriculum developers in the teachers' guide assigned to teaching third year secondary school textbook entitled *"New Prospects"* or inherited in the literature of language proficiency and communicative competence. The test framework guided us to identify the cognitive constructs in specifying the process dimension (the cognitive skills of knowing and doing) and above all, the test format.

In the Algerian context of education, the framework is embodied by three elements of the competency-based approach as stated in the ministerial documents accompanying the textbook and the teachers' guide (Ministry of National Education, 2017). This approach is formally applied in Algerian schools where it focuses on performance, authenticity and problem solving. Its building blocks stem around the *constructivist* views to learning by doing (active learning) and Bloom's *Taxonomy of Educational Objectives* that delineate constructing one's own meaning of the world and cognitive levels of knowledge, comprehension, application, analysis and synthesis. This insight directed us to emphasize communicative real word performance and constructing one' own meaning rather than recalling knowledge, attempting to link instruction to the outside word reflecting the real world needs of learners by creating realistic contextual situations for language use and designing challenging tasks stimulating problem solving and critical thinking skills.

Moreover, when considering the theory underpinning the present study, we also refer to the literature in the domain of language proficiency, and Bachman's approach to communicative language testing in particular. Bachman (1990), a leading figure in language testing, suggests an insightful approach to define language proficiency. It is called the *real life approach* which views language proficiency as being so "characteristic of the performance of competent language users" (p.41). The relevant characteristics for measuring language proficiency contain almost all those traits representing language use or contextual features like the relationship existing between interlocutors, precise content areas and situations, and aspects of the language system

(grammar, vocabulary, and pronunciation). For reasons of space, expansion of the approach is discarded; we just emphasize what feeds the current research purpose and context.

Given this insight, it follows that language/lexical proficiency, or what Bachman refers to as language ability, is the research construct that includes the domain of language use that requires the design of real life situations, and inclusion of one component of the language system that of vocabulary knowledge, which is believed necessary for discourse production. Accordingly, performance, in this study, will sample students' use of target words in real life situations to communicate about historical, ethical, educational and social topics prescribed by the textbook content.

What is further suggested is that the defining principles of language ability drawn from Bachman's model are to some extent considered relevant to defining lexical ability, the core issue in this study. For the research purposes, we also opted for the Bachman's approach to language proficiency, because it approaches language ability from authenticity perspective and this is particularly typical to performance testing. On the other hand, it specifies attributes like language use (precisely word use in our case), specific content areas and situations (the targeted textbook content is of topical or thematic orientation), and aspects of language system namely vocabulary. These attributes, we believe, are representative of the valid measures of our construct of lexical ability.

One more argument for the actual testing conceptual framework adheres to the communicative approach to language testing, an approach which is evidently influenced by Dell Hymes communicative competence. The trend seems to be reflected in the following assertions stated by Alduais (2012):

- "learners are assessed with the use of performance tests on the basis of communication acts they perform be it receptive or productive, and social roles must be integrated in any test" (p.206).

- "(…) language is a means of communication and the major aim of tests is testing the learners' ability to communicate effectively; that is to use productively what they have learned receptively (prospectively)" (p.206).

In connection with the above quotes and, as a starting move, we might offer the following proposal:

- Designing a vocabulary test based on performance tasks;

- The ultimate goal of these tasks is to gauge communicative ability; targeting learners' ability to communicate effectively using prior knowledge of lexis;

- Emphasizing productive language use (vocabulary use);

- Authenticity should be reflected in item types; designing authentic tasks setting out real world situations;

- Integrating social roles, such as asking leaners to act like a school counselor giving advice to students suffering from exam stress and anxiety, or to act like a representative in the UNECEF meeting standing against child labor, roles such as these appear in the eight communicative tasks constructed for the study; and

- Emphasizing language use to transfer receptive vocabulary knowledge to productive vocabulary use in contexts.

As can be noticed, the theory severed a foundation stone for this test study design and development.

The aforementioned section briefly hinted at the theory behind the current study and henceforth it defines the construct operationally. The second phase of designing performance allows the test developer to practically link the research construct to the observed performance or behavior, i.e., word knowledge. It also contributes to address test construct validity. In this section, we specify students' levels of lexical ability that would be reflected in the test content. Because, tests are said to be "the operational definition of the construct" (Bachman, 1990, p.46). We refer to our construct as one that involves that mental ability represented in the form of test scores by depicting learners' lexical or vocabulary ability. Based on Ntaion's (2001, 2013) dimensional approach to vocabulary testing, we would design a test that requires test takers to recognize target words, build their word families, and above all use them in novel real world situations. These attributes, or word knowledge aspects, based on the learning outcomes mentioned in the textbook (see Appendix B), we assume would elicit lexical performance because

in achievement tests, “the learning objectives of the syllabus construct the theoretical definition of the ability to be tested” (Bachman, 1990, p.46).

4.5.2.2. Test Specifications

Test specifications, also known as the blue print (Genesee & Upshur, 1996) guided us to the design of items and the various tasks for the present type of assessment. Test specifications consulted aid test developers to make their assessment consistent with the content, including skills, abilities and knowledge of the subject matter (Johnson et al., 2009). Alignment with textbook unit outcomes was a starting point for test specifications that determines the content of the tasks.

In the literature of language assessment, three components should appear in the description of the design process: “test framework”, which is discussed before, “test specifications” and “specifications table” that will be explored next. Test framework, according to Johnson et al. (2009), delineates the construct or domain of measurement that, in turn, has two dimensions of content and cognition. In our case, the construct is the lexical knowledge and ability that we expect students to be able to reveal after a course of instruction in a written form. Vocabulary knowledge entails word meaning recognition, word formation processes, word use and contextualization. The content dimension is selected with reference to these thematic orientations: Ancient Civilizations, Education, Ethics in Business, and Feelings and Emotions as prescribed in the textbook *New Prospects* (see section 4.5.2.4.2, p. 219 for a comprehensive description of this coursebook).

The cognitive dimensions of vocabulary assessment lie on: the conceptual understanding of the content subject area namely lexical items, knowing and doing of these words involving application of knowledge in word combination and word use, and practical reasoning where students are expected to imagine the situation and delineate it along words that fit. The selection of test content and activities will depend on the fields of vocabulary, the nature of words; the constituents of knowing and doing words, and vocabulary themes. Test specifications, on the other hand allude to

> A detailed description for a test that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, and scoring rubrics and

> procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices. (AERA, APA, & NCME, 1999, p.183)

From this comprehensive extract, it follows that test specifications as a second step in designing performance determines the skeleton of the test and the testing conditions such as the rating procedures surrounding it, and extends to the willingness to explore the psychometric properties affecting test results.

In sum, the test writer will adopt a number of divergent standards in designing the current performance assessment; beginning with writing a table of specifications, defining content, expected responses, then developing scoring rubrics and validation procedures.

4.5.2.3. Table of specifications

A table of specifications *"*lists the test content and specifies the number or percentage of test items that cover each content area" (Johnson et al., 2009, p.28). The table of specifications set for this research project is embedded in the test specifications, which involves content standards, content dimension and process dimension, the distribution of items by subject and/or skills, item type, scoring, and timing. These are summed up in Table 4.1 (p.182).

Content standards for the current assessment are constructed with reference to the "statements of what students should know and be able to do" (Johnson et al., 2009, p.37) as a result of instruction. These statements are drawn from *"New Prospects'"* textbook outcomes listed at the outset of every unit's sequences (See Appendix B). We selected those outcomes sounding relevant to the assessment of the construct of vocabulary knowledge. Content standards are extracted from the defined textbook, because it represents a source material used to teach the participants before joining their course in the Department of English at ENSB. It seems useful to recall that, this current attainment test aims to measure first year university students' (pre-service teachers') productive vocabulary knowledge based on the prior knowledge they brought to the class. The content standards gauge whether these students' will be able to use the tatget

Content standards	Content dimension	Process dimension	Scoring	
-demonstrate an understanding/conceptualization of words (meaning recognition of both the situation problem and the target words). - apply knowledge of word formation processes (affixation). - use the target words in novel situational context (doing). - from a cognitive perspective, students should be able to make connection between familiar words and the new situation (analysis and synthesis).	Fields of vocabulary (themes, word forms, word meaning, spelling and word use)	It includes cognitive skills of knowing and doing: use known words in new real-world application; practical reasoning)	Holistic scoring rubrics	
Item type	Number of items	Points per item	Total points per items	Time allocated per each task
Performance tasks	08	**5** per teach ask	40	15 minutes per each task

Table 4.1: Table of Specifications

words that were taught in the third year secondary school textbook appropriately in new contexts with the right word meanings and word forms. Whether they can demonstrate their potential to contextualize words adequately in given written discourse settings is the major research concern. In this context, communicative/lexical competence equals knowledge and skills necessary to complete the tasks; that is, how far students are competent to use target words productively and more appropriately in various communicative situations.

Because we opted for an embedded comprehensive context dependent vocabulary measure, it is also worthy to mention that among the language content standards or secondary education curriculum standards, especially set for BAC students, are the writing skills. The English language writing standards set for terminal classes include genres or types of writing whereby students can distinguish the features characterizing each. It is highly arguable to assume that students who are expected to produce these

kinds of texts are able to write for a variety of purposes, and those used to write essays, whether argumentative or expository …etc. are also expected to reveal their ability to write short paragraphs or create a variety of responses, but the question that might be posed is whether examinees can use target words productively to generate these targeted responses or not. The content standard discussed here is not listed in the above specifications table because the optimal research objective is not to assess the writing proficiency but lexical competence, and writing is a mode via which vocabulary can be assessed. Issues on mechanics, use of cohesive devices, and organization of ideas are put in a second order activity and are slightly considered when weighting the examinees' performances.

Up to this point, three processes have been considered, and now developing assessment requires considering a number of standards.

4.5.2.4. Elements of Test Specifications (Specifications for test writers)

In the literature of educational measurement, developing assessment content requires considering a number of standards: determine content, population to be tested, developing items and tasks, administering examination, scoring and reporting, and formally assessing the psychometric characteristics of the test to institute change if necessary. The standards followed in the process of test design and development are further explored along the upcoming sections.

4.5.2.4.1. Defining Content

Content specifications for educational measurements are often equated with content standards (Johnson et al., 2009, p.37). Content standards, in this sense, denote the expectation that students in third year secondary education will know or do as a result of instruction; these standards are clearly stated in the textbook "*new prospects'"* content. They are represented in the learning outcomes prescribed by the textbook content and syllabus. A standard related to vocabulary is students' ability to recognize meanings of words, apply word-formation processes to create novel words, and use familiar words in simulated authentic contexts. We specify competence in vocabulary by this ability to problem-solve, reason and communicate in written form. Accordingly, the emphasis of the assessment should reflect the subject matter content; that is learners'

vocabulary knowledge or lexical competence introduced in the language learning outcomes (word building) stated overtly by the textbook (see Appendix B).

In the light of the above statements, we feel it sensible to provide a workable description of the textbook in paly with more emphasis on content standards.

4.5.2.4.2. Textbook Description in Relation to Content Specifications

At the start, *New Prospects*, as described by the Ministry of National Education (2017), seeks to improve three competencies of interaction, interpretation and production. These competencies are expected to be mastered by secondary education third year students in relation to specific tasks and situations. These tasks and situations basically deal with language systems, namely, syntax, morphology, vocabulary, pronunciation, and spelling through six graded units.

In reference to the textbook units, we designed eight vocabulary complex communicative tasks corresponding to the three competences of interaction, interpretation and production of various written but not spoken messages elicited via paper and pencil test format. These competences are taken into account while designing the test as they form the basis of the course as described in the teachers' guide (Ministry of National Education, 2006). They are integrated in the item types structure and development corresponding to the textbook prescription:

- Interaction with the tasks refers to the participants' expressions and responses to the tasks situations using appropriate vocabulary corresponding to particular communicative situations.

- Interpretation is also considered in the test design as it reflects learners' ability to comprehend the tasks prompts, contexts and instructions, transmit their ideas and provide responses accordingly.

- As for production, the respondents are required to produce messages as they are expected to produce written messages using the different types of written discourse (narrative, descriptive, argumentative, expository, and injunctive) corresponding to a given communicative situation. These tasks are problem solving situations/tasks that

require learners to solve problems that seem to be more or less complex to create written production (Ministry of National education, 2017).

The optimal objective of integration, or the exit profile, to be attained by third year secondary school students after three years of secondary education learning English is stated clearly by the Ministry of National Education (2006):

> Dans une situation de communication, et sur la base d'un support oral ou écrit, l'élève doit produire un message écrit d'une vingtaine de lignes, dans un type de discours écrit choisi (descriptif, narratif, argumentatif, expositif, injonctif), correctement et lisiblement. (p.5)

The textbook *"New Prospects"* also assumes that it focuses on the knowledge that improves a given use of English. The test, therefore, takes word use in particular contexts as one criterion in the alternative assessment. The content or themes of the test items are selected only in the four compulsory units of: *Ancient civilizations; Ethics in business; Education in the World: comparing educational systems*; and *Feelings and emotions* as cited respectively in "*New Prospects"*. Note that tasks, test items and item types are used interchangeably throughout this thesis.

It is worth mentioning, here, that only the four units target words and themes are considered in the study. Almost all the respondents belonged to the literary streams in secondary education, to whom the four above unit contents were exposed, and the remaining units, however, are taught to scientific streams; topics of *Astronomy*; *Food and Health*) among which we had one participant, whose exam paper was discarded. Because, among the four units the student missed the topics of *education in the world* and *feelings and emotions.* Alternatively, the informant was exposed to the defined scientific topics together with Ancient *Civilizations* and *Ethics in Business* those shared with literary streams.

With reference to the third-year secondary school program and performance/competence assessment literature, we designed a test composed of eight complex situations that were meant to elicit vocabulary knowledge, mainly word recognition, word-formation, word use and ability to interpret messages and construct well-structured paragraphs. Performance/competence-based assessment requires an

investigation of students' performance by means of various complex tasks. The idea is supported by Shavelson et al. (1993) who state that multiple tasks and variety of methods are required to judge and triangulate on students' performance. Inspired by this, the current assessment consists of a variety of tasks designed within performance-based paradigm because performance tasks allow researchers and teachers to collect evidence not only about students' knowledge and the content domain but also on what s/he can do with that knowledge (Darling-Hammond & Adamson, 2010).

In keeping with the aforementioned content standards prescribed by the Ministry of National Education (2017), we put more emphasis on the last three competencies of interaction, interpretation and production in the assessment of student's performance. For content specifications, we focused on morphology, vocabulary, and spelling henceforth, whereas syntactical aspects of language were excluded because the research focus is put on content/lexical words rather than functional or grammatical words, but this does not mean that issues of accuracy are totally excluded from assessment purposes. Additionally, since this test is a pencil and paper test, pronunciation or sound system which is an oral performance would be eliminated from the content of the test. Hence, automaticity of free language production, one way of vocabulary assessment, is entirely excluded from our research design. In brief, in this performance test, the students' responses should show comprehension and production of language setting under contextual constrains. Because performance testing for Bailey (1998) is connected with the performance of given professions or a set of contextualized communicative functions.

In a nutshell, performance/competence-based assessments need a set of tasks sampled from a universe of admissible observations. Accordingly, we attempted to develop a group of open-ended tasks to gauge the students' performance eventhough it is quite difficult, needing more effort and skills, to design similar forms of tasks to investigate lexical performance variability. Because task sampling variability is so important in validating performance task scores (Huang, 2012), it is important to vary task items in terms of themes (see section 4.5.2.4.3. below) to depict the participants' vocabulary knowledge.

Recalling the discussions in the second chapter devoted to a survey of the literature seems useful as that chapter highlighted different key steps in vocabulary assessment. What words to test and how to test them is rather a key step towards an effective test design. Selection of words to test is an issue that is considered next.

4.5.2.4.3. Word Selection

The vocabulary test will assess knowledge of word meanings, word formation processes and use of particular content words encompassing nouns, verbs, and adjectives, but not knowledge of function words as they carry syntactical function although they represent the most frequent words in the English language. Eventhough "more frequent words should have priority in learning and assessment" (Read, 2012, p.258), content words will serve the sample of assessment content. Because content words are central to the focus of the textbook **"***New prospects"*. In its introduction, it stresses, "the emphasis is on a thematic orientation" (Ministry of National Education, 2017, p.4). The syllabus translated into the textbook, as it sounded, is topic or content-based and this type of syllabus emphasizes vocabulary instruction and learning. In this regard, words to be assessed are selected from the textbook content. The issue has been much clarified by Read (2012) who points out that conventional vocabulary tests select words from worldwide famous word lists whether integrated in the textbook or not, they are rather meant to be given to students to study. This means that it is possible to sample words from the textbook instead of worldwide corpuses to serve our research purposes.

The present test contains 24 target words purposefully selected from the prepared word list drawn from the intended textbook corresponding to the tasks situations that are thematically-geared. Though the words included in all the tasks exceed this number (54 words), the test takers were restricted to use only 24 lexical items in their performance, as suggested by the experts participated in the test validation process that will be further explored in detail in the subsequent sections of this chapter. We did not randomize the selection and presentation of these target words to enable an estimation of the words that the learner deeply knows out of the words in the word list. During the word selection process, we emphasized only the words that belong to the four units devoted to teaching Foreign Languages stream, because students of this stream

represent the majority of the baccalaureate holders that are given priority to be recruited in ENSB. The words to be tested are assumed to be part of the respondents' receptive knowledge, because they were taught during their previous course of instruction, i.e. being taught throughout their secondary school third year, and drawn from the textbook '*New Prospects''* thematic units, these stimulus words are, seemingly, expected to be part of the students' productive use.

Selecting some words for the testing purposes from the textbook was to ensure an equal opportunity of vocabulary mastery. Selecting words from readymade word lists was not possible as "to prevent the ceiling effect". The test covers only the words existed in the textbook that is those encountered during classroom instruction being based on the teaching/leaning textbook. The target words involve nouns, verbs and adjectives chosen correspondingly to the intended instructional outcomes. Developing morphological/word classes knowledge is favored in each textbook unit.

Because vocabulary includes a large corpus of words, it is not possible to test all the words in the textbook. Nation (2013) suggests a solution for this asserting that, "a good vocabulary test has plenty of items – around 30 is probably a minimum for a reliable test" (p. 536-537). To our context, they are reduced to 24 words because they are intended to be largely contextualized within written discourse from the part of students that may fit the different communicative problem solving tasks. Unlike those multiple choice tests that Nation talked about, that seek to estimate examinees' vocabulary size where samples of words are large and the minimum is 30 words, this test is meant to measure the students' quality of vocabulary knowledge emphasizing productive vocabulary knowledge in particular.

The above emerging step sounds very much in conformity with Schmitt's (2010) steps used in developing vocabulary tests examined in Chapter 1. This step includes what words to test, but before tackling how these words are invested in the overall task design and structuring, we opt firstly for describing the test takers, pursuing Genesee's and Upshur (1996) framework then we move straitforwards to developing test items.

4.5.2.5. Specifications for Test Users

This section provides a description of the examinee's characteristics. As mentioned already in section 4.4 (p. 172), the test takers are newly registered in the first year university level, they are students recruited in the Department of English at ENSB, their age ranges between 17 and 18 years old. As to the gender, only three male test takers participated in the study, the remaining are females representing 100 of the entire sample. The English language is positioned as the second FL. The current test of vocabulary knowledge is a test designed to assess the English language proficiency of students who do not have English as their first language those joined the postgraduate college where English is the medium of instruction. The sample represents a selected group that seems to be superior in terms of language ability. It is, thus, treated as homogeneous due to the fact that the baccalaureate holders have the same exist profile and similar academic testing conditions.

The aim of this test is to elicit the students' lexical ability they come up with to university. When they come to university their entrance profile is apparently determined by their exposure to English language for seven years. The exit profile, meaning the optimal objective of integration, of third year secondary school students has been described in the teachers' guide (Ministry of National Education, 2017) as follows: in a communicative situation, based on an oral or written prompt, the students can produce a correct readable written message of twenty lines corresponding to a given type of discourse (descriptive, narrative, argumentative, expository and injunctive). Supposedly, the participants in the sample are able to create a well-written production in relation to a given text or communicative situation.

The test takers were considered in the content and instructions addressed. They were informed of the scoring criteria so that they can discriminate between the tasks from the view point of their relative importance. The same can hold true for clarity of instructions and timing. These students were administered one test where time allocated for task completion is not that important because the focal point is performance not speed. One hundred and thirteen EFL students at ENSB participated in the study. It is important to note that some papers were not considered for reasons of task completion.

One might question why first year university students and not other levels? The suggested answer would be that measuring competencies in higher education is considered a neglected aspect in the Algerian context of education, especially for the baccalaureate holders newly recruited in ENSB, the teacher training school. Stress is put on the exist profile that learners bring to the English language classroom to further build programs fitting their own background level and knowledge. To this end a number of tasks have been constructed.

4.5.3. Developing and Describing Items and Tasks

One fundamental step perused in the test item types design and development was to unpack the performance expectation; we identified the competencies intended for measurement compliant with the third-year secondary education program, vocabulary knowledge literature together with performance-based instruction principles i.e., the quality criteria used in planning and writing the tasks. On that account, a number of concerns have been taken into consideration in the current assessment when designing performance tasks. These concerns have been developed and synthesized from the reviewed literature (Johnson et al., 2009):

a. Identifying new, rich and authentic situations: we produced original tasks to ensure that the students have never been exposed to and we provided them with rich realistic/real contexts where they could invest their previously acquired knowledge and skills, thus bridging instruction to the real world needs of learners.

b. Developing prompts: tasks contain context, task input and instruction.

c. Taking into account social and professional life and real world needs of learners that have been contextualized within the test items content.

d. Respecting students' cognitive and educational levels by designing tasks of suitable level of difficulty.

e. Paying attention to instruction clarity and accurateness to avoid ambiguity and openness to various interpretations.

f. Drafting, piloting, validating, editing and revising the tasks: the tasks were presented to educational specialists including university teachers specialized in educational

psychology and assessment, secondary schools teachers in charge of teaching third year level as well as inspectors belonging to secondary education. In addition to piloting the tasks on a sample of students.

g. Creating scoring rubrics for judges to assess students' performance and training raters to effectively use the scoring rubrics.

Besides these concerns, a checklist of content validity, which has been constructed to approve whether certain parameters are so characterizing the tasks or not. This checklist is composed of 21 parameters (see Appendix C) distributed to specialists, university/secondary education teachers and inspectors, to confirm tasks validity and enlighten us with more comments and suggestions for fruitful task revisions.

We have thus far been in a position to make the skeleton or overall structure for our test. The section coming next will further describe the process of generating test items.

4.5.3.1. Developing Task Structure

Once the competency, which is specifically linked to lexical ability of particular words, has been delineated and the target words have also been sampled, we relied on the expertise of a group of teachers and inspectors working in secondary schools and university professors in the design of assessment tasks. As a result, eight communicative complex tasks have been constructed and divided into four-category format. Each category is thematically-geared, and is grouped under one of the four headings in accordance with the four textbook units labels: the first two tasks are around the topic of *Exploring the Past (Ancient Civilizations)*; task three and four explore the theme of *Ill Gotten Gains Never Prosper (Ethics in Business)*; tasks five and task six illustrate *Education in the World (Comparing Educational Systems)*; and the last two tasks represent vocabularies associated with *We Are A Family (Feelings and Emotions)*. As for grading and sequencing, the tasks are thematically-ordered according to the order of the textbook units themes.

Based on the review of literature on performnce assessment (see Chapter 2), we took into consideration the conditions that sound necessary for the design of task format

and content. These conditions have been evidenced by the defined expertise, and hence the results can be assumed to be representative of students' performnce:

- The tasks are new and have never been encountered by students previously;

- The tasks integrate knowledge, skills, and attitudes;

- They require complex performance involving knowledge, skills and attitude and complex instructions;

- The tasks are appropriate to the students' educational level and of appropriate level of difficulty;

-They integrate interesting topics and cultural content appropriate for students;

- They contain clear prompts, topics and instructions, and even the target words were previously covered;

-They contain common knowledge and familiar topics so as respondents can generate their responses;

As stated above, eliciting whether these features (parameters) are reflected in the test items, a checklist of content validity for teachers was assigned to be filled in a pilot study and after in a pre-administering phase of the test (see Appendix C), and a questionnaire for learners was distributed to support test items quality (Appendix E).

In the writing process, we also relied on a "To-Do List for Constructing Performance Tasks" from "*Assessing Performance: Designing, Scoring, and Validating Performance Tasks"* by Johnson et al. (2009, see pp. 92-93); and Design features of writing prompts for large-scale assessment (p.104). These two procedures guided the development of the tasks in terms of topic familiarity, task context, cueing the expected performance, prompt format/ response expected format, scoring criteria, prompt clarity, guidance towards what to include, how to structure the response, and indication of how responses will be evaluated, and finally and most importantly, emphasizing the alignment of the prompts to the curriculum and content standards.

It is worth mentioning that when writing the tasks we favored authentic situations to serve as stimulus materials. Baron (1991) comments that performance activities

should be drawn from real-world contexts in order to actively involve students so as they learn that there is a real purpose for learning and that their knowledge and skills are worth outside the classroom.

The overall design of our assessment is composed of a set of performance tasks. The response format is neither multiple choice nor extended response (essay), but the format is a short constructed response; short performance tasks or short paragraphs.We will use the same form for all the tasks; that is, in accordance with the same specification table, including content and skills, but with various situation problems and different themes. The process is named *Parallel Forms* in the terminology of performance assessment. We opted for this form for the many reasons that are stated by Johnson et .al (2009):

> Administration of parallel forms reduces the possibility that any change in scores is attributable to examinee familiarity with the test. Parallel forms also serve a security purpose in credentialing examinations and other high-stakes tests in that no one test is readministered every time, making it more difficult for examinees to learn the exact items on the test. (p.42)

The authors further comment that tasks with parallel forms of an assessment adhere to the same specifications table as they address similar content and processes, but are constructed with different problems. In our case, all the test items are designed in relation to the same content standards, content dimension, and process dimension as shown in Table 4.1 (p. 182) as well as process or cognitive dimension. However, they incorporate variations in situation problems.

We also tried to construct novel tasks situationalized in new contexts so we can stimulate the students' cognitive construct. Because if students have already been exposed to the same situation problem then no mental processes, like analysis and evaluation, will have an effect, and their real abilities will not be shown, rather only memory and recognition will stand for this. Once testing techniques resemble activities implemented for learning, Bachman (1990) asserted that the potential negative bias of the test method would be minimal, because learners are expected to do the same tasks. Thus, we attempted to design the same task format but with different situations, topics, instructions, and target words in a thematic order.

4.5.3.2. Classifying Assessment Tasks

Performance-based assessment requires a variety of tasks to illustrate students' performance, that is why we depended on different complex tasks in formulating the overall test design. This alternative assessment contains eight communicative tasks (situations of integration) organized in four sections each of which has two tasks that are, as mentioned earlier, thematically linked. How the tasks are sequenced is indicated as under:

- Tasks one and two stem from the topic of Ancient Civilization;

- Tasks three and four cover the topic of Education in the World;

- Tasks five and six involve the topic of Ethics in Business; and

- Tasks seven and eight derive from the topic of Feelings and Emotions.

Tasks content/prompts together with their instructions and model responses are to be described next.

4.5.3.3. Item Types Presentations with Instructions and Sample Answers

With the help of professors, teachers and inspectors, we designed eight open-ended tasks corresponding to the targeted competency organized under four type formats analogous with the topics of civilization, ethics in business, education in the world, and feelings and emotions. Each task includes a simple prompt as a stimulus material. Each prompt specifies only one familiar topic and what to do with it, describing for instance, and the related target words henceforth. The prompts set realistic communicative situations where a learner can be asked by his teacher or someone else to do a task (e.g., talk about one ancient civilization). Each task illustrates how it has a social and professional goal being daily experienced by students. These open-ended tasks require learners to be creatively, critically, mindedly involved in operating thinking when doing the tasks. The tasks content and structure together with their expended responses are presented underneath.

Task 1: *(Time allocated 15minutes)*

You have been told a story by your grandfather about ancient Egyptian/Sumerian/Ottoman civilization that he has read about. You wanted to share the story with your classmates!

Recite (describe) one civilization in which you report imaginary or real events (or feature of the civilization), ***using*** *the* ***verbs*** *of the following words:* ***flourishment, invention, rise along, fall into ruins.***

Expected Response:

Ancient Egyptian civilization is among the oldest civilizations in the world. It ***rose along*** *the Nile River. It* ***flourished*** *during periods of peace. Ancient Egyptian* ***invented*** *agriculture (irrigation in particular), but afterwards it* ***fell into ruins*** *because of wars.*

Task 2: *(time allocated 15 minutes)*

You listened to a radio documentary about Algerian public revolutions in different regions just after 1830's fighting against French colonization led by famous Algerian leaders.

You have been questioned by your teacher of history to give a short account of the characteristics underlying ancient Algerian leaders.

Use the ***adjectives*** *of these words to best describe those heroes or their civilization:* ***knowledge, peace, nomad, war.***

Expected Response:

Algerian leaders of public revolution (or revolutionary leaders) were ***knowledgeable*** *eventhough they were not exposed to formal education. They used to be* ***peaceful*** *and lived a* ***nomadic*** *life before 1830, but because of the French invasion, they became* ***warlike*** *to free their lands.*

Task 3: *(time allocated 15 minutes)*

You have been introduced to a new pupil (or a face booker) unfamiliar with your school who is coming from other English countries.

Describe your classroom or school to him/her in terms of location or curriculum studies ***using*** *four (04)* ***adjectives*** *of these words:* ***situation, compulsiveness, attendance, graduation, education, qualification, assessment, and training.***

Expected Response 1:

Our school named (x) is ***situated*** *along a beautiful river. Having this picturesque view is of no importance if not accompanied by fruitful efforts. It provides a* ***compulsory*** *and obligatory education. Students* ***attending*** *the school for three years will be* ***assessed*** *in the Baccalaureate examinations and those who succeed will be* ***qualified/graduated*** *for further studies.*

Expected response 2:

The school ***educational*** *system is very high. It is based on three major regulations: students' attendance is* ***compulsory****, their academic achievement is effectively* ***assessed*** *and higher* ***qualified*** *teachers provide enough training for students for better qualification.*

Task 4: *(time allocated 15 minutes)*

Imagine that you are an experienced teacher and you are asked to plan an ideal school and present it in a conference attended by UNICEF members in an international meeting.

Decide which sort of school it would be.

Use *the* ***adjectives*** *of the following words to describe your plan:* ***discipline, educate, construct, train.***

Expected Response:

Our school should take an effective ***disciplinary*** *education to enhance our* ***educational*** *system. We need* ***constructive*** *training and teaching. Through instruction we should inherit social awareness, skills and autonomy to provide well enough* ***trained*** *individuals to function adequately in society.*

Task 5: *(time allocated 15 minutes)*

Suppose that you were a businessman working on mobile phones company. You suffer from many counterfeiters infringing (imitating) your copyright material. Remind them that imitating property is theft and can cause great deal of financial loss.

Choose four (04) words you think will support your ideas. ***Use*** *the* ***"ing"*** *form or the noun of these words:* ***counterfeit, defraud steel, consume, deceive, and advertise.***

Expected Response:

Imitating property is ***theft****. Many* ***consumers****, especially those with low income are obliged to buy your* ***counterfeits*** *to fulfill their needs eventhough they know that imitated property is just like stolen property. Counterfeiting aims at* ***deceiving/defrauding*** *people because you are* ***advertising*** *products of low quality and lack safety standards. Eventhough brands are too expensive, but their goods are ethical and do not intend to defraud. These brands spend large amounts of money advertising against* ***counterfeiting*** *so counterfeiting causes financial loss.*

Task 6: *(time allocated 15 minutes)*

Imagine you are Human Rights activist against children labor. You have been asked to make an appeal to address people about the need to eradicate this malpractice.

Select ***four*** *(04) words from the list and* ***use*** *their* ***adjectives*** *to address your audience:* ***unethicality, ethics, crime (organization), boycott, violence,*** *and* ***exploitation.***

Expected Response:

We need to ensure that employers or managers of companies are above child labor if so their companies are considered ***criminal*** *organizations. By this age children are supposed to attend classes at schools but not to be* ***exploited*** *by employers! It is* ***unethical*** *to deprive children from schooling as it is illegal to practice* ***violent*** *behaviors against children to increase productivity. Companies of this kind should be* ***boycotted****.*

Task 7: *(time allocated 15 minutes)*

Suppose that you were a school psychologist, and you are asked to lecture (deliver) a conference in which you give a piece of advice to a student suffering from stress because of the Baccalaureate examinations.

Choose four (04) words stated underneath to achieve success and relief; transforming the words to ***adjectives: self-satisfaction, humour, optimism, worry, fun, stress.***

Expected Response:

The Baccalaureate examinations are a nightmare almost for all students. But if you follow these recommendations you will inevitably succeed: First, be ***humorous*** *because being so has good effects on your health and social behavior and your studies henceforth. Second be* ***optimistic*** *and always think that you can be among the winners. Third, do never be* ***worried,*** *BAC examinations are just like normal exams. Finally, whenever you feel* ***stressed*** *and you did your work harder try to shift your attention towards other activities like reading a book or watch a film.*

Task 8: *(time allocated 15 minutes)*

You have been told a comic, or tragic story by a friend and you also want to tell him/her a story or a scene you watched in a film or read in a book about a recent comic or tragic story.

Describe one scene that deeply affected you and express your feeling ***using*** *four adjectives (04) of these words:* ***sadness, fear, grief, anger, dislike, funny, irritated, relaxed, happy, crying, proud, satisfied.***

Expected Response:

It was a dark night when the actor was ***irritated*** *by the noise. When he went down the stairs carrying out a candle to check the source of the noise, he was* ***fearful****. Down the stairs there were two brightened red eyes; it was a cat that made him* ***angry****. He finally felt* ***relaxed.***

Before accomplishment of the above performance tasks format and structure, they were tried out along various pilot studies using a number of measures.

4. 6. Pilot Studies

Decisions about item types prompts, content, and structure were not perfected until supported by experts in the subject area; a number of university professors, secondary school teachers and inspectors participated in the pilot study. These

participants were asked to respond to a checklist of items containing the eight performance tasks and a set of evaluative criteria (parameters). Relying on the subject area experts is believed to be useful to ensure the quality of content, its representativeness and the scoring system.

In practice, the research work went through four pilot study stages that will be explored in details afterwards:

Phase 1: *Piloting the first draft checklist of content validity containing the tasks eight prompts for further item writing*

Phase 2: *Reviewing the psychometric properties and credentialing the test content and format afterwards, by means of administering a final version of the checklist of content validity of the performance tasks to the expertise;*

Phase 3: *Piloting the test for students to elicit their performance, validating its content and trying out the scoring guide; and*

Phase 4: *Administering a survey questionnaire (see Appendix E) for learners to check their attitudes towards test item types complexity, novelty, instruction clarity, cognitive load, authenticity, …etc. in order to further refine the tasks structure when necessary by means of reviewing the psychometric properties.*

Phase 1: Piloting Checklist Item Writing and Task Format

The checklist pilot study was carried out in three weeks. Ten first draft checklists of content validity of the test performance tasks were distributed to ten teachers between, 16thApril till 7thMay, 2017, but only seven checklists were collected and filled in by three university professors and four secondary school teachers. It is worth noting that my supervisor's effect on the validation process as an expert in the assessment domain cannot be overlooked. Accordingly, many adjustments related to the tasks format including timing, topics, instruction and some other parameters modifications had been made based on the participants' comments and suggestions.

As far as time allotment is concerned, teachers suggested lengthening it to 15 minutes instead of 10 minutes per each task. Two teachers suggested that students, at

this level, need five minutes at least to understand the topic then the remaining ten minutes for interpretation and production.

As for topics, only one topic was eliminated from task eight; it was that of love that was said to be sensitive to the students' culture as three teachers stated. It was replaced by the topic of stress and anxiety within the BAC context. As such, this task content does not satisfy bias and sensitivity issues.

From the perspective of task format and structure, no comments had been made on the prompts and contexts. However, task questions were criticized by two teachers specialized in performance assessment for being so long and specifying task response format. In other words, they suggested not to ask overtly students to write a paragraph, or constitute sentences or even to ask them to write a short memo as the emphasis is put on assessing vocabulary not writing. Accordingly, instructions had been changed to either describe or write a short account, or make an appeal, etc.

Based on the expertise's comments, a number of changes had been made at the level of checklist content validity parameters; some were deleted and others adjusted. The whole procedure is described as follows:

- Parameter 3 *"The task is a real life-like activity (reflects real life language use)"* sounds like parameter 4 *"Task context is linked to the social and professional life of students"* that is why the later condition was deleted.

- We also integrated parameters 2 and 5 into one condition (parameter 2 *"The task includes knowledge, skills and attitudes acquired by students previously"*; parameter 5 *"The task integrates different skills, attitudes rather than a simple recall"* by keeping the latter as it is.

- Parameters 3 and 12 were considered alike that is why the twelfth was deleted and the third was kept as it is (parameter 3 *"The task is a real life-like activity (reflects real life language use*); 12. *The task encourages students to actively use the language").*

- Parameters1*"The task addresses positive values and goals relevant to the society and education system"* and 22 *"The task (topic and content) is meaningful to the students"*

were integrated and replaced by *"The task (topic and content) addresses positive values and goals relevant to the students, society and education systems."*

In general, the number of parameters was reduced from 24 to 21 by adding one parameter linked to time efficiency to the checklist as teachers insisted on it. These were all the modifications occurred otherwise, the tasks characteristics and conditions were said to be good and suffice the conditions necessary for judging the performance tasks and almost all the parameters were available across all the tasks.

Phase 2: Reviewing the Psychometric Properties Based on the Checklist

To enhance credibility on the tasks format and prompts construct and content validity and in order to review the psychometric properties of the assessment tasks, a pilot study was conducted in three phases. The tasks were validated by the teachers' expertise in phase 2 and by learner's performance and learners' survey questionnaire in phases 3 and 4 respectively. This practical aspect is accomplished through the distribution of 50 checklists to teachers (but only 22 returned them) whom, we believe, are sufficiently acquainted with knowledge about the learners' needs, level, exist profile, content standards, and the quality of tasks performed in their classrooms as well as the whole approach applied in secondary education for teaching and assessment. Furthermore, these teachers also have this ability to convert content of subject area into tasks, items or instructions, being aware of the content that best represents the domain of assessment.

The expertise helped in the content and construct validation process. Twenty two teachers responded to the checklist of content validity including the list of performance tasks and a set of evaluative parameters (see Appendix C). In fact, a pilot study or what is referred to in the literature of performance assessment as practice analysis (Johnson et al., 2009) was conducted for credentialing the test. This checklist contained different items to weigh the importance of items and tasks in terms of structural adequacy (content coverage, prompts, situations and context, instructions), appropriateness to learners' level, needs and schemata, …etc. Each task was supplemented by a set of evaluative criteria that may include authenticity, importance, relevance to the course, complexity, frequency of occurrence i.e., how often the task is performed in class, and

the criterion of criticality depicting whether the task content is harmful to the students identity and culture …, etc.

In practice, the checklist on the test content and format was devoted to the teachers with definite distributed parameters. Reviewing the quality of the performance tasks needs the use of the constructed evaluative checklist (see Appendix C), composed of 21 items we adapted from Herman et al., (1992), Perlman (2002), Johnson et al. (2006), and Genesee & Upshur (1996). For reasons of tasks content validity, we relied on judges estimates. This is a convention in psychological and educational measurement, because content validity is a key criterion in performance/competency assessment. Furthermore, to check the degree to which the tasks designed reflect the necessary parameters, or the quality criteria, for competency/performance assessment programs and complex task performance, we built this checklist of content validity, which is necessarily available in any complex task, and calculated concordance coefficients between the 22 judges to ensure that all the tasks meet the necessary conditions. In assigning scores to the responses provided, we used the following rubric or numerical ratings to score the building blocks of this checklist:

0 = this element is not evident in the resource being evaluated (0 is the equivalent of No); the number 0 indicates that the task does not satisfy each parameter in the checklist.

1 = this element is very evident in the resource being evaluated (1 is the equivalent of Yes).

We estimated the concordance coefficient by dividing the number of agreement between participants in each parameter on the number of agreements and disagreements between the judges.

The pilot study participants are acquainted with knowledge about the learners' knowledge, abilities, needs, level, interest, profile and the like. Among the sample, there were three inspectors, 16 secondary school teachers, and three university professors interested in competency assessment. After the pilot study had been conducted, calculating concordance coefficients for each communicative situation have been

achieved, the results obtained are presented in the following four tables containing all the parameters and dichotomously included the task number.

Construct Validity of Task one and Task two

Number	Parameter	Concordance Coefficient	
		Task 1	**Task 2**
1	The task matches the objective (measuring student's vocabulary knowledge).	0.95	0.90
2	The task integrates knowledge, skills and attitudes rather than a simple recall.	0.85	0.9
3	The task is a real life-like activity (reflects real life language use.	**0.63**	**0.71**
4	The task stimulates students thinking skills.	**0.77**	**0.73**
5	The task is intended for assessing students' performance.	**0.78**	**0.7**
6	The task allows students to respond differently.	0.85	0.80
7	The task contains the necessary information needed to arrive at the correct answer.	**0.61**	0.80
8	The task is structured so that students can have control over response format.	**0.77**	0.90
9	The task structure does not favour some students at the expense of others.	0.86	0.90
10	The task is of appropriate level of difficulty for students.	**0.59**	**0.66**
11	The task (topic and content) addresses positive values and goals relevant to the students, society and education systems.	0.86	0.90
12	Task content is not sensitive (not hurting or displeasing).	0.81	0.85
13	The task includes cultural content appropriate to the students.	0.90	1.00
14	Task content allows students to know how the response will be evaluated.	**0.59**	**0.7**
15	Task instruction is clear.	0.95	0.9
16	The content of the task is understandable.	0.90	0.90
17	The task is familiar (practiced previously) but its content is new to all students.	**0.71**	**0.75**
18	The task is interesting to the students.	**0.66**	0.80
19	Task situation contains the basic components (context, prompt, and instruction).	0.95	1.00
20	The language of the task is free from lexical redundancy.	0.8	0.90
21	Time allocated for the task is sufficient.	**0.36**	**0.47**

Table 4.2: Concordance Coefficients between Experts' Judgments on Task 1 and Task 2

As shown in Table 4.2, most of the concordance coefficients are beyond (0.80) in most of the parameters of task one and two, consequently their content validity is high except for some parameters that have low concordance coefficients. These are taken into consideration via modifications based on the participants' suggestions.

Two teachers think that task one is not authentic since it says that a student might not be told a story by his grandfather because he is adult enough and can not stay on his knees and listen to a story. They also think that the topic of old history is not realistic. Even so, the topic of ancient civilization is still considered as it serves to be part of the first mandatory unit of the targeted textbook.

Construct Validity of Task three and Task four

Number	Parameter	Concordance Coefficient	
		Task 3	**Task 4**
1	The task matches the objective (measuring student's vocabulary knowledge).	1.00	0.95
2	The task integrates knowledge, skills and attitudes rather than a simple recall.	0.90	1.00
3	The task is a real life-like activity (reflects real life language use.	0.86	**0.7**
4	The task stimulates students thinking skills.	0.81	1.00
5	The task is intended for assessing students' performance.	0.90	0.9
6	The task allows students to respond differently.	0.86	0.89
7	The task contains the necessary information needed to arrive at the correct answer.	0.85	**0.78**
8	The task is structured so that students can have control over response format.	0.90	**0.61**
9	The task structure does not favour some students at the expense of others.	0.90	0.85
10	The task is of appropriate level of difficulty for students.	0.81	**0.75**
11	The task (topic and content) addresses positive values and goals relevant to the students, society and education systems.	0.81	0.85
12	Task content is not sensitive (not hurting or displeasing).	0.90	0.85
13	The task includes cultural content appropriate to the students.	0.80	0.90
14	Task content allows students to know how the response will be evaluated.	**0.65**	**0.6**
15	Task instruction is clear.	0.95	0.90
16	The content of the task is understandable.	0.95	1.00
17	The task is familiar (practiced previously) but its content is new to all students.	**0.71**	**0.8**
18	The task is interesting to the students.	**0.66**	0.90
Number	**Parameter**	**Concordace Coefficient**	
		Task 3	**Task 4**
20	The language of the task is free from lexical redundancy.	0.95	1.00
21	Time allocated for the task is sufficient.	**0.5**	**0.42**

Table 4.3: Concordance Coefficients between Experts' Judgments on Task 3 and Task 4

As Table 4.3 illustrates, most of the tasks meet the parameters; that is content validity is high since most of the concordance coefficients are above (0.80) some parameters, however, seem to be low but even so we relied on the parameters and teachers' recommendations to write the final draft.

Construct Validity of Task five and Task six

Number	Parameter	Concordance Coefficient	
		Task 5	**Task 6**
1	The task matches the objective (measuring student's vocabulary knowledge).	1.00	0.90
2	The task integrates knowledge, skills and attitudes rather than a simple recall.	0.90	0.85
3	The task is a real life-like activity (reflects real life language use.	0.81	0.85
4	The task stimulates students thinking skills.	0.90	0.90
5	The task is intended for assessing students' performance.	0.80	0.95
6	The task allows students to respond differently.	0.95	0.95
7	The task contains the necessary information needed to arrive at the correct answer.	0.90	0.95
8	The task is structured so that students can have control over response format.	0.81	**0.76**
9	The task structure does not favour some students at the expense of others.	0.95	0.8
10	The task is of appropriate level of difficulty for students.	**0.77**	0.80
11	The task (topic and content) addresses positive values and goals relevant to the students, society and education systems.	0.90	0.95
12	Task content is not sensitive (not hurting or displeasing).	0.95	0.90
13	The task includes cultural content appropriate to the students.	0.95	0.95
14	Task content allows students to know how the response will be evaluated.	0.80	0.80
15	Task instruction is clear.	1.00	1.00
16	The content of the task is understandable.	1.00	1.00
17	The task is familiar (practiced previously) but its content is new to all students.	0.80	0.85
18	The task is interesting to the students.	**0.71**	0.9
19	Task situation contains the basic components (context, prompt, and instruction).	1.00	0.95
Number	**Parameter**	**Concordance Coefficient**	
		Task 5	**Task 6**
20	The language of the task is free from lexical redundancy.	0.85	0.85
21	Time allocated for the task is sufficient.	**0.5**	**0.57**

Table 4.4: Concordance Coefficients between Experts' Judgments on Task 5 and Task 6

Table 4.4 reveals that most of the concordance coefficients between the participants are very high in tasks five and six. Thus, content validity of these tasks is very high (mostly above 0.80). The remaining low coefficients show that topics are not interesting and that time is not sufficient. These limitations would be refined.

Construct Validity of Task seven and Task eight

Number	Parameter	Concordance Coefficient	
		Task 7	**Task 8**
1	The task matches the objective (measuring student's vocabulary knowledge).	0.90	0.95
2	The task integrates knowledge, skills and attitudes rather than a simple recall.	0.90	0.86
3	The task is a real life-like activity (reflects real life language use.	0.90	0.86
4	The task stimulates students thinking skills.	0.86	0.86
5	The task is intended for assessing students' performance.	0.85	0.86
6	The task allows students to respond differently.	0.86	1.00
7	The task contains the necessary information needed to arrive at the correct answer.	**0.76**	0.94
8	The task is structured so that students can have control over response format.	0.80	0.85
9	The task structure does not favour some students at the expense of others.	0.85	0.8
10	The task is of appropriate level of difficulty for students.	**0.76**	0.86
11	The task (topic and content) addresses positive values and goals relevant to the students, society and education systems.	0.86	0.95
12	Task content is not sensitive (not hurting or displeasing).	0.86	0.86
13	The task includes cultural content appropriate to the students.	0.85	0.90
14	Task content allows students to know how the response will be evaluated.	**0.66**	**0.61**
15	Task instruction is clear.	1.00	0.95
16	The content of the task is understandable.	0.95	0.95
17	The task is familiar (practiced previously) but its content is new to all students.	**0.71**	**0.76**
18	The task is interesting to the students.	0.90	0.85

Number	**Parameter**	**Concordace Coefficient**	
		Task 7	**Task 8**
20	The language of the task is free from lexical redundancy.	0.90	0.86
21	Time allocated for the task is sufficient.	**0.54**	**0.5**

Table 4.5: Concordance Coefficients between Experts' Judgments on Task 7 and Task 8

As indicated in the table above, most of the concordance coefficients are beyond (0.80) in most of the parameters of tasks seven and eight, consequently their content validity is high except for some parameters that have low concordance coefficients. These are taken into consideration via modifications based on the participants' suggestions.

Phase 3: Test Piloting for Learners

For reasons of test validation, the checklist of content validity (see Appendix C) was produced, and thus distributed to the expertise for more refinement and validity. In the meanwhile, the tasks were also piloted in the first week of October 2017 to 50 first year EFL university students at ENSB. Based on the expertise suggestions, one hour and half time was allocated, however, some students could not complete the last two tasks. The students were asked to respond to the tasks and feel free to supply their comments on the tasks in the blanks left after each task when necessary.

Based on the students' comments and suggestions, a number of adjustments and amendments were made to refine the test structure. These are described along the following:

- A maximum time would be provided (to the total time of one hour and half) in the next administration and the word *'pedagogue'* in task four was replaced by *'teacher'* as it sounds difficult for some learners;

- Instructions were separated from the prompt and were written in italics and the target words were written in bold to stress them because some students found it somehow hard to recognize the tasks instructions;

-Most of the students felt frustrated of being restricted to particular four words in each task (in a controlled productive vocabulary test) and, therefore, they proposed to add some many words to the list to shed some freedom on their performance; and

- Space left for the answer was enlarged from four spaced lines to six lines based on the learners' suggestions and the target words were increased to six too. For reasons of space and timing limitations, accompanied by other test structure and content shortcomings, the test was piloted again for further refinements.

Phase 4: Piloting for further Validation and Trying out the Scoring Guide

According to Nation (2001), issues like reliability, validity, practicality and washback should be regarded in the process of designing vocabulary tests. To this end, a group of 33 participants (but only 30 papers were considered) coming mostly from Algiers, Bedjaia, Blida, Media, Tizi Ouzou, etc. took part in the validation study on April 16th, 2018. The pilot study was intended to be conducted in late October, 2017 with other groups, yet it was only delayed and not piloted up to this time because of the students' long strike that lasted for about 5 months. They were studying at different high schools and got their BAC, which means they have the same exit profile. A maximum time was allocated to complete the eight performance tasks. Two raters scored their performance "to attain more reliable scores and fairer judgments, literature suggests including two or more raters in the assessment process and assessing students' writing ability through various tasks or topics" (Lee et al., 2002, as cited in Sari & Han, 2022, p. 41). The results thus produced are displayed along the upcoming G analyses tables:

Facet	Label	Levels	Univ
Persons	P	30	INF
Tasks	T	8	INF
Raters	R	2	INF

Table 4.6: Observation and Estimation Designs

Source	SS	Df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	189.917	29	6.549	0.326	0.326	0.326	34.7	0.105
T	26.633	7	3.805	0.038	0.038	0.038	4.1	0.031
R	4.033	1	4.033	0.010	0.010	0.010	1.0	0.014
PT	135.617	203	0.668	0.213	0.213	0.213	22.7	0.035
PR	26.217	29	0.904	0.083	0.083	0.083	8.8	0.029
TR	7.633	7	1.090	0.028	0.028	0.028	3.0	0.017
PTR	49.117	203	0.242	0.242	0.242	0.242	25.7	0.024
Total	439.167	479					100%	

Table 4.7: G Study Table (Analysis of Variance)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% Relative	Absolute error variance	% Absolute
P	0.326		.....		.....	
	.....	T	.....		0.005	5.0
	.....	R	.....		0.005	5.0
	.....	PT	0.027	32.0	0.027	28.2
	.....	PR	0.041	49.8	0.041	43.8
	.....	TR	.....		0.002	1.9
	.....	PTR	0.015	18.2	0.015	16.0
Sum of variances	0.326		0.083	100%	0.094	100%
Standard Deviation	0.571		Relative SE: 0.288			Absolute SE: 0.307
Coef_G relative	0.80					
Coef_G absolute	0.78					

Table 5.8: Estimated Variance Components and Reliability Coefficients

Note. Grand mean for levels used: 2.292, Variance error of the mean for levels used: 0.025, Standard error of the grand mean: 0.158.

In general, what can be interpreted from the above table is that the test being piloted produced reliable scores, as they reached an acceptable level of reliability 0.80 and 0.78. Yet, the raters still need more training in the use of the scoring rubrics because "The use of highly trained raters and monitoring procedures helps to reduce the random error and bias introduced by human raters" Schmidgall (2017, p.2). View sections 4.7 (p. 213) and 4.9 (p. 217) for a full description of the study scoring rubrics development and practice. Here, the sources of variance are not that important to explore in details. That is why only the reliability index was considered.

Phase 5: Reviewing Psychometric Properties based on the Survey Questionnaire

As it has already stated, a survey questionnaire was also distributed to leaners after phase 4, after the test had been administered to them, the aim of which was to provide more refinement, when necessary, on test structure and content. The survey

questionnaire contains 16 items constructed based on the literature review, 10 questions per each task and 6 questions on the whole test. Items number is not exclusive to avoid bothering the 33 students with a lot of questions as they were required to answer the items just after the test had been piloted. The items revolve around task difficulty, words complexity, task novelty, cognitive stimulation or complexity, tasks authenticity, interest, meaningfulness, opportunity to practice previously acquired knowledge, instruction clarity, assessing knowledge, skills and attitudes acquired before. The questions are close-ended with yes/no format intended to obtain more objective results. After it had been distributed to the participants, concordance coefficients were calculated by means of counting the number of agreement between the participants in each question on the number of agreements and disagreements between them. The results thus produced are shown in Tables 4.9, 4.10, and 4.11.

Number	**Question items**	Concordance Coefficient		
		Task 1	**Task 2**	**Task 3**
1	Is the task difficult?	**0.15**	**0.24**	**0.18**
2	Are the words of the task complex?	**0.15**	**0.15**	**0.39**
3	Is this problem situation new for you?	**0.33**	**0.39**	**0.18**
4	Does the task stimulate your thinking? (challenging)	0.75	0.87	0.75
5	Is the task realistic?	0.93	0.84	0.90
6	Is the task meaningful?	0.93	0.96	0.93
7	Is the task interesting?	0.90	0.90	0.81
8	Does the task encourage you practice the language you acquired previously?	0.87	0.72	0.87
9	Is the instruction clear?	0.93	0.81	0.78
Number	**Question items**	**Concordance Coefficient**		
		Task 1	**Task 2**	**Task 3**
10	Does the content of the task reflect what you have learned in 3rd year secondary school (knowledge, skills, and attitudes)?	0.87	**0.33**	0.78

Table 4.9: Concordance Coefficients between Students' Attitudes on Tasks 1, 2, and 3

Table 4.9 indicates that most of the concordance coefficients between the participants are very high in tasks 1, 2, and 3. Subsequently, content validity of these tasks is high (mostly above 0.70) they are closer to 1.00. The remaining low coefficients show that few students agreed that the tasks are difficult and they contain complex words, and that the problem is not new for them. The respondents might consider the tasks as being not

new because the context resembles those they have in class and the topics and target words are relevant to their course of instruction. As concerns topics and time, the students showed that topics are not interesting and that time is not sufficient. Time would be lengthened, and topics however are not modified as the test targets the students' lexical competence based on their prior knowledge which is relevant to the defined program.

Number	Question items	Concordance Coefficient		
		Task 4	**Task 5**	**Task 6**
1	Is the task difficult?	**0.33**	**0.21**	**0.12**
2	Are the words of the task complex?	**0.18**	**0.24**	**0.12**
3	Is this problem situation new for you?	**0.30**	**0.18**	**0.12**
4	Does the task stimulate your thinking? (challenging)	0.78	0.72	0.84
5	Is the task realistic?	0.90	0.93	0.93
6	Is the task meaningful?	0.96	1.00	1.00
7	Is the task interesting?	0.93	0.87	0.96
8	Does the task encourage you practice the language you acquired previously?	0.84	0.90	0.96
9	Is the instruction clear?	0.93	0.90	0.90
10	Does the content of the task reflect what you have learned in 3rd year secondary school (knowledge, skills, and attitudes)?	**0.30**	0.96	0.93

4.10: Concordance Coefficients between Students Attitudes on Tasks 4, 5, and 6.

Similar to Table 4.9 above, Table 4.10 shows low concordance coefficients for Q1, Q2 and Q3 which means that only a minority of respondents agreed that Tasks 4, 5, and 6 are difficult, and contain difficult words which are not new. Otherwise, all the concordance coefficients vary from acceptable to very high except for Task 4 that has low concordance coefficients for Q 10 meaning that the task does not reflect what was taught; but when shared with the expertise they refuted this result.

Compared to the two previous tables, the following table also displays similar results as concerns Q1, Q2, and Q 3 in terms of task difficulty, target words complexity and task novelty. The remaining concordance coefficients range between acceptable, high, and extremely high (from 0.60 to 1.00).

Number	Question items	Concordance Coefficient	
		Task 7	**Task 8**
1	Is the task difficult?	**0.09**	**0.12**
2	Are the words of the task complex?	**0.12**	**0.21**
3	Is this problem situation new for you?	**0.18**	**0.42**
4	Does the task stimulate your thinking? (challenging)	0.75	0.84
5	Is the task realistic?	0.93	0.87
6	Is the task meaningful?	1.00	0.87
7	Is the task interesting?	0.93	0.81
8	Does the task encourage you practice the language you acquired previously?	0.81	0.75
9	Is the instruction clear?	0.96	0.93
10	Does the content of the task reflect what you have learned in 3rd year secondary school (knowledge, skills, and attitudes)?	0.78	0.60

Table 4.11: Concordance Coefficients between Students Attitudes on Tasks 7 and 8

As to the students' attitudes towards the test overall questions that were posed to explore their attitudes towards the six test conditions like time and space devoted for test completion, students' ability/skill to respond appropriately in a written format, topic and content coverage, fairness in weighting performance, Table 4.12 that comes next will display the results generated on the whole test:

Number	Question items	Concordance Coefficient
11	Is time allocated sufficient for task completion?	0.45
12	Is space left sufficient?	0.45
13	Can you have control over response format?	0.69
14	Can you determine the topic of the task?	0.93
15	Does the test provide lots of tasks (different ways) for you to show that you understand what was taught?	0.81
16	Do some tasks deserve more points than others ?	0.60

Table 4.12: Concordance Coefficients between Students' Attitudes on the Whole Test

Table 4.12, again, exhibits some limitations related to time and space sufficiency as some students responded (Q11 and Q12 resulted in 0.45 concordance coefficients). The project group suggested a solution to this issue by adding half an hour in the final test administering, eventhough they indicated that these learners must have low language ability as time is sufficient for them to complete the tasks. Additionally, space was also

extended to eight lines instead of six. The remaining concordance coefficients are either acceptable, high or extremely high denoting an acceptable level of test validity.

By now, issues of content and construct validity have been considered, the subsequent sections are to be devoted to developing a scoring guide to quantify the students' performances and to further examine the current test reliability index.

4.7. Development of Scoring Rubrics and Quantifying Observation

Simultaneously to the administration phase, we developed a scoring guide whereby scorers can evaluate students' performances. The starting point for the development of the study rubrics was by aligning the scoring criteria with the content standards already established for the current domain of interest. In essence, to ensure the reliability factor in our test it is important to assign scales to quantify the observed behavior, the lexical/language performance. Bachman (1990) states that one way to determine measurement attributes of language is by defining levels of performance or language ability on a scale. Here we talk of scoring rubrics. By scoring is meant to award marks on the observed lexical level of performance.

After the performance tasks had been constructed and administered, we reviewed some literature on assessing performance and language assessment to first select appropriate scoring rubrics and then write the rubrics appropriate to the present testing purposes. Our decision fell into the use of holistic rubrics with every performance task. To facilitate the scoring process, we established scoring schemes including a description of performance levels or criteria.

This step is so reciprocal to achieve test validity and reliability, and to score students' performance on the intended tasks. It is universally agreed that selecting adequate rating scales facilitates the process of scoring the responses and increases the levels of agreement among raters. We opted for the holistic rubrics to score the students' performance in every task in the test, so a single scoring rubric was relevant to the study parallel forms. The holistic scales assign distinctive scores along the various criteria. Holistic scorings are favored because the main emphasis of the recurring theme is drawing an overall picture of participants' vocabulary knowledge not diagnosing vocabulary or writing weaknesses. This means that the focal point is on the optimal

lexical performance not its parts. In this context, analytic rubrics are totally discarded since the purpose of the assessment is not diagnostic in principle, it is rather summative which is meant to determine the students' lexical competence. The criteria, therefore, are to be used globally without analyzing responses traits separately. We used the holistic rubrics to determine the students' performance as one entity using five point scale where no response as a criterion was considered as null unless learners exhibit no attempt on a certain task performance. These rubrics are described along Table 4.13.

Criteria	Grading	Description
More Advanced (out of 5)	**100%**	Student demonstrates: - An excellent response with an absolute understanding /conceptualization of the words (and topic). - Ability to appropriately use all target words in novel context with good word choice. - Successfully uses affixation to form new words. - Writes a well developed paragraph with relevant information and well-structured sentences with perfect grammar.
Advanced (out of 4)	**75%**	Student shows a skillful response with: - Complete and correct recognition of words and word use with some specific word choice. - Correct use of word building processes. - Writes a clear paragraph with some development, including some relevant information. - Exhibits control over sentence boundaries; errors in grammar, spelling, and punctuation
Proficient (out of 3)	**50%**	Student's response shows a partially complete and correct of: meaning recognition, word use, affixation (either missing an important detail or including a small incorrect detail). - writes a clear paragraph with little development; has few information - has simple sentences and simple word choice; may exhibit uneven control over sentence boundaries. - Has sentences that consist mostly of complete, clear, distinct thoughts; errors in grammar.
Nearly proficient (out of 2)	**25%**	Student's response indicates unsatisfactory or incorrect word use, or only one correct use.
Novice (out of 0 or1)	**1%** **0%**	No word in the instruction would appear, but s/he describes something related to the topic. The value zero (0) shows the absence of the attribute.

Table 4.13: Holistic Scoring Rubrics Template and Marking Schemes

According to Jhonson et al. (2009), this type of performance can be assessed with rubrics, because it is a constructed response and represents a work that demonstrates understanding of concepts/ words. Besides, the authors state that it can be evaluated via holistic rubrics because "Holistic rubrics often follow a paragraph or block format" (p. 157), and because these rubrics give importance to the whole performance and view performance as greater than the sum of its parts. Based on this conceptualization, the responses follow a paragraph format that entails the targeted vocabulary knowledge, including sentence boundaries, ideas coherence, cohesion, etc. and other mechanics shown in the scoring system template implemented in the research work (See table 4.13, p.214).

Establishing valid criteria for assessment is a complex process. Yet, after reviewing the related literature the process becomes much more possible. The scoring rubric, as displayed in the table above, describes five levels of performance. Under each performance level, there are a set of criteria set for evaluation and assigning scores. We used five specific proficiency-based rubrics to smoothen the scoring stage. The evaluative criteria in this research work revolve around the important components of productive vocabulary knowledge. The holistic rubrics template adapted from Mertler's (n.d); Penny et al.'s (2000); Huang's (2009); and Smit's & Birri (2014) scoring guides, was constructed with relevance to the current demands. We assigned 5 points (represents 100%) for more advanced responses; 4 points (75 %) for advanced answers; 3 points (50%) for proficient performance; 2 points (25%) for nearly proficient response exhibition; 1 point for novice students (1%) featured out no target word use but demonstrating something related to the topic of the task. The value zero (0) shows the absence of the attribute or intended ability. The scoring levels or levels of performance are described from high to low in Table 4.13 (p.214).

From the scoring rubric template, it follows that the targeted performance response should show comprehension and production of language under contextual constrains. Because performance testing for Bailey (1998) is connected with the performance of a certain professions or a number of contextualized communicative functions.

After the scoring guide had been developed and then tried out in scoring the piloted performance, the in-depth productive vocabulary test was administered to the intended sample.

4.8. Test Administration

After the development of the eight performance tasks had been accomplished, the assessment was administered to the test takers. To achieve a certain level of standardization and uniformity in test administration, we provided the necessary instructions and training to the staff (ENSB teachers) as regards timing, duration and overall testing procedures. Because, in tests administered "to assess the examinee's knowledge, skills, or abilities, standardization helps to ensure that all examinees have the same opportunity to demonstrate their competencies" (AERA, APA, & NCME, 1999, p. 61).

Accordingly, the same test was administered to all the participants at the same time from 11h: 00 to 13h: 00 on Tuesday 27th November, 2018. Note that it was impossible to administer it before this date as an entrance test because of a long strike that started right from the outset of the academic year. Invigilators distributed the exam papers that were provided to them before the examination time. They read the questions and monitored the students. Every student worked individually and silently read from the two printed pages containing the eight tasks and they wrote their responses within the same exam sheet. No reference materials had been provided as more stress is put on testing vocabulary. The students were also deprived from using MP3 players, digital watches, cell phones, pagers and other tools that might serve them with source materials.

During the administration process, the invigilators arranged the seating with space left between the test takers to avoid cheating. Besides, the staff was acknowledged of the type of assistance they might provide for students in case they rise questions (e.g., response format and clarifying instruction) and, then, gathered the completed performances. These assessment conditions were considerd to ensure fairness and comparability of the resulting test scores. For test security purposes, we had also visited the examination rooms and we were ready to answer any question that might be posed

by the examinees. Finally, we collected and signed in all the packaged test materials at the end of the assessment.

We, together with the head of department of English language, organized the staff, arranged schedules and rooms, and oversaw the exam sheet distribution. The test invigilators distributed the test papers and read the instructions to the test takers. They were also responsible for creating a healthy examination atmosphere; preventing the examinees from making noise, talking or sharing answers. These standardized conditions and uniformity of directions set for the invigilators and examinees were meant to decrease the impact of extraneous factors, external to examinees, on the students' observed scores because score variability ought to be resulting from individual differences and abilities and not from test measures and varying conditions (Clemans, 1971).

4.9. Perfromance Scoring Procedures

All the eight written communicative tasks responses were scored by two secondary school teachers, one of which is also an inspector. They were opted for being acquainted with knowledge about the learners' exit profile, skills, and abilities. The teachers were trained on the use of predetermined rating scales; both the tasks and the scoring system were so far explained. Beyond this, the raters scored some sample tasks, and exam sheets henceforth, until they felt ready to score the overall performance tasks. Working in a project team, during the training session, a holistic rating system was emphasized whereby the raters were informed to assign one score per each level of performance to which the test takers had accomplished/met the expected goal of the assignment. The judges were asked neither to score specific traits of the attribute of vocabulary knowledge nor to assign multiple scores. They were informed to put primarily more emphasis on the productive vocabulary knowledge construct in communicative settings depicting their fluency, and not writing proficiency although the tasks were in a written context dependent measure (see Read, 2000) and this of course does not mean that some aspects of writings were almost ignored.

The raters were trained in the scoring process and were informed to put more focus on vocabulary knowledge; word meaning, word form and word use. Since vocabulary

is tested via writing, some writing conventions should be considered and others should not. Thus, the extent to which the examinee exhibits control over sentence formation, usage, and mechanics were valued. Nevertheless, these were not heavily weighted in determining the overall lexical competence/performance. Issues on sentence formation and boundaries involve correctness, complexity, meaning clarity, ending punctuation. As to usage, the raters were also informed to consider subject–verb agreement, possessives, etc. Mechanics include capitalization, spelling, within paragraph punctuation and sentence breaks.

We developed the scales of benchmarks and then edited and confirmed by a project group of teachers, containing three secondary school teachers among which we had an inspector, and two university teachers, one is specialized in psychological and educational measurement and the other in EFL assessment, who is the researcher's supervisor. The scoring rubrics reflect the levels of vocabulary knowledge attribute served to covey a standard reference for the raters to avoid any scoring biases, or undesirable effects to ensure consistency among the resulting ratings to achieve inter-rater consistency.

Now that the procedures used for data collection have been examined, the subsequent sections will be devoted to sketching out the procedures used for data analysis.

4.10. Methods of Data collection and Analysis: Applying G Theory

After the collection of the data from the vocabulary performance test, the results were analyzed quantitatively using G theory principles and procedures. In this research, the data analysis goes through the application of G theory phases and the various decisions made on the measurement facets, and the study designs; whether crossed or nested designs. The decisions also involve conducting G and D studies relevant to each research question. Finally and mostly important, employing the EduG software package to insert the data or the observed scores and analyze them statistically.

When conducting a G study it is important to identify the facets of measurement that might affect validity and reliability of measurements. This section describes the three assessment conditions relevant to the study endeavors through different stages

conducted in G study analysis and D studies. We applied the principles of G theory in planning G analyses and study designs, identifying facet sampling status, through four stages: observation design, estimation design, measurement design, design evaluation and optimization design.

4.10.1. Observation Design

The first step in planning a G study is to identify the facets deployed in the measurement process. In this study, the investigated sources of measurement error include tasks *(t)*, raters *(r)*, and themes *(h)*. These are referred to in G theory as facets of measurement (instrumentation facets) with students considered as object of measurement (differentiation facet). The errors linked to these facets are assumed to be relevant to the assessment of lexical competency. In performance/competence-based assessment, students' performance is distinct to the extent that it necessitates a relatively increased number of *raters* to obtain reliable scores. This is because it depends on open-ended tasks that require a complex illustrative performance from the part of the learner. Observation and judgment of students' performance in one occasion is not sufficient to determine their competency in a particular domain. Although occasion is considered important in testing reliability measures in G theory, in this study, however, we opted for administering a test only in one occasion to give examinees an even equal test condition for test performance and because of the teaching load constraints, we could not administer the test twice to all first year ENSB students. For time and space management reasons, we could not consider occasion as one more source of variance in this research work. Testing vocabulary did not occur on more than one occasion.

When considering other performance/competence-based assessment characteristics, three facets were investigated: 1) a variety of tasks were designed to further cover the intended content of the course. Task variety, in this context, corresponds to content coverage; 2) two raters were involved in the scoring process; and 3) an examination of the students' performance across different themes in addition to estimating their performance variance between tasks and task items. The themes were incorporated as a facet of measurement to check the extent to which they are attributable

to measurement error and can affect reliability of scores and whether examinees' performance varies across the four themes that are relevant to the textbook graded units.

As for raters, two secondary education teachers were observed and made independent judgment concerning students' performance using specific scoring rubrics. Raters participated in the present study are secondary school teachers, and particularly teaching terminal classes. They are believed to be useful to provide accurate observed scores; being acquainted with some familiarity with the students being assessed and the scoring procedures. They are also aware of the content and lexis being exposed to these students throughout "*New Prosects'* " textbook units. Each rater scored each assessment task individually after being introduced to and trained on the use of the rating rubrics. The aim of varying the raters is to see the extent to which the use of different raters to weigh students' performnces would contribute to measurement error and affect score generalizability.

The aforementioned facets of measurement, we believe, might result in variance in students' performance across (a) tasks because of variance in the tasks themselves; (b) variance in raters' estimations; and (c) variance in the stability of scores across themes of each task. Accordingly, we conducted different G study designs, crossed and nested, that take into consideration the different facets of measurement that have the potential to affect students'observed scores and assessment precision. These are tasks, raters, and themes.

To answer the research questions, this study relied on an examination of reliability and validity of test scores obtained from an assessment of lexical competence using G theory. The latter helped us determine the variance components that can affect generalization of scores in addition to eliciting the best methods that can improve measurement procedures to obtain more generalizable scores. To this end, throughout the study, we opted for a suitable measurement design that sounds consisitent with the literature of G theory. Within its stages, we explore the research study data gathering designs status.

4.10.2. Data Collection Designs

The facets of interest in this research are tasks *(t)*, raters *(r)* and themes *(h)*. To estimate their relative impact on score generalizability and dependability, we used three G study designs: one is fully crossed and the two others are either partially nested or patially crossed faceted designs. Hence, the current study is a generalizability study with one crossed and two mixed designs as the assessment conditions are often influenced by task variability, rater variance as well as themes variance.

Here in the observation design stage, it is worthy to consider facets and their interrelationships in the overall G study design. These facets can be either crossed with one another or partially nested. The first study is a two-faceted crossed design **(p × t × r)** because all the students performed all and the same tasks and all the raters scored all the same tasks or every students' performance across parallel forms; i.e., tasks with similar response frormats. The crossing relationship is applied to three facets of persons, tasks and raters, with students as objects of measurement (differentiation facet) and raters and tasks as instrumentation facets. This generalizability design was set forward to estimate reliability of scores obtained from a performance assessment. In general, all G study designs were meant to ensure dependability and validity of students' scores obtained from a set of complex communicative tasks.

1. Fully Crossed Design with Two Facet Universes: this G study design was used as a procedure for the data collection.It uses three crossed facets of measurement: student by task by rater. These are infinite random crossed facets and the students are treated as an object of measurement or as a distinctive facto; whereas, the raters and tasks are said to be sources of variance or instrumentation facets. In the following table (Table 4.14) are the facets, their symbols, levels and their various observations:

Facets	Facet abbreviation	Observations	Facet levels
Students	**P**	From student 1 to 113	113
Tasks	**T**	From task 1 to 2	08
		From task 1 to 8	
Raters	**R**	From rater 1 to 2	02

Table 4.14: Observation Design with Two Facet Universes (P×T×R)

Note. Levels: different conditions of measurement, Facet: test items, raters, and themes, Observation: observed behaviors/measurements

The crossing relationship existing between facets of measurements is further presented in Figure 4.1, where one facet of person is crossed with two other facets of task and rater denoted as **(p × t × r)** the two facet universes.

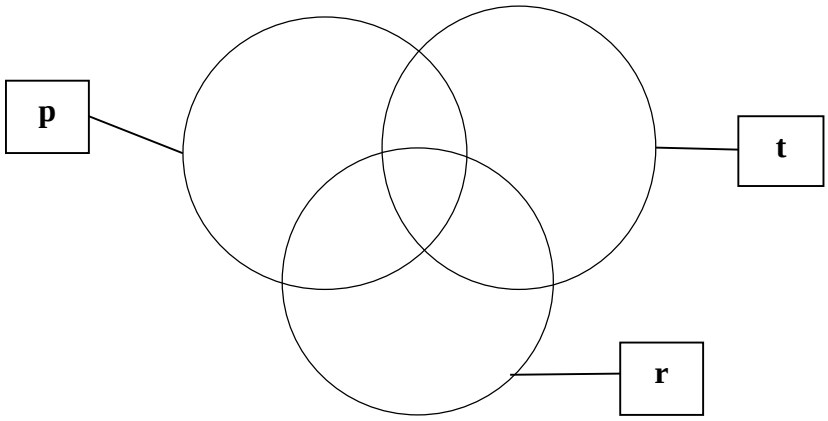


Figure 4.1: Variance Partition Diagram for the Observation Design (P×T×R)

2. Partially Nested Design with Two Facet Universes: the second G study design symbolized as **P(T:H)** incorporates Persons *(P)*, Themes *(H)*, and Tasks *(T)*. When considering inter-relationships existing between facets of measurement, the current G study design is composed of one facet crossed with two nested facets. Despite the fact that all the 113 participants had performed the test 08 tasks, and hence for the themes underpinning the tasks, all the three facets are partially nested. In this case, the three facets are expected to be fully crossed but the logistics justifies the nesting interrelationships and at the same time the crossing interrelationship is also approved.

This means that the 08 tasks are categorized under 04 theme headings, each two tasks are nested within one theme (all the tasks are nested within the themes); each two tasks in the G study are embedded within one theme. Task1 and Task 2 under the theme of *Ancient civilizations*, Tasks 3 and 4 consider the theme of *Education*, Tasks 5 and 6 tackle *Ethics in Business*, and Tasks 7 and 8 deal with *Feelings and Emotions*. The abbreviation for the facet interrelationship 'T:H' denotes that each 02 tasks from the 08 tasks are nested within one theme. In Table 4.15 under, are the facets, their levels and abbreviations together with their corresponding observations.

Facets	Facet abbreviation	Observations	Facet levels
Students	**P**	From student 1 to 113	113
Tasks	**T**	Form task 1 to 8	08
Raters	**R**	From rater 1 to 2	02
Themes	**H**	From theme 1 to 4	04

Table 4.15: Observation Design P(T:H) with One Facet Crossed with Two Nested Facet Universes

In the following figure (Figure 4.2), the crossing and nesting facet interrelationships set for the partially netsed design are displayed, where the facet of person is crossed with two nested facets of task and rater denoted as **p×(t:h)**.

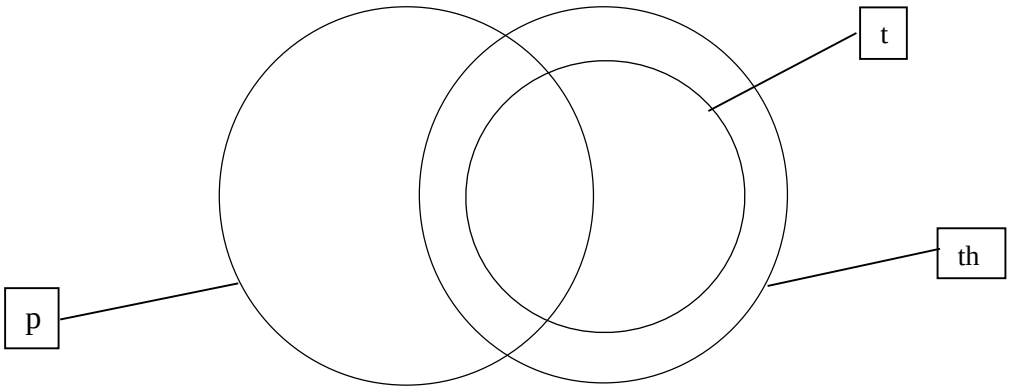


Figure 4.2: **Variance Partition Diagram for the Observation Design P×(T:H)**

3. Partially-Crossed P×R(T:H) Design with Three Facets: when compared to the first and the second data collection G study designs, the observation design in this

model changes in terms of facet interrelationships and facet levels. The three facet p×r(t:h) partially-crossed design encompasses four facets that characterize the G study. These are persons *(p)*, raters *(r)*, themes *(h)* and tasks *(t)*. The three facets of persons, raters and tasks are crossed because all the students performed all the item types (tasks) and all the raters weighted every students' performance, and hence they rated the same set of tasks. Each two tasks of the 08 tasks are classified under one theme of the four suggested themes discussed above, the facet of tasks is nested within the facet of themes, and the observation design is therefore PR(T:H) or (T:H)PR. The facets levels and their observations for this design are summed up in Table 4.16 and in Figure 4.3.

Facets	Facet abbreviation	Observations	Levels
Students	**P**	From student 1 to 113	113
Tasks	**T**	From task 1 to 8	8
Raters	**R**	From rater 1 to 2	2
Themes	**H**	From theme 1 to 4	4

Table 4.16: Observation Design with Three Facets Design P×R(T:Th)

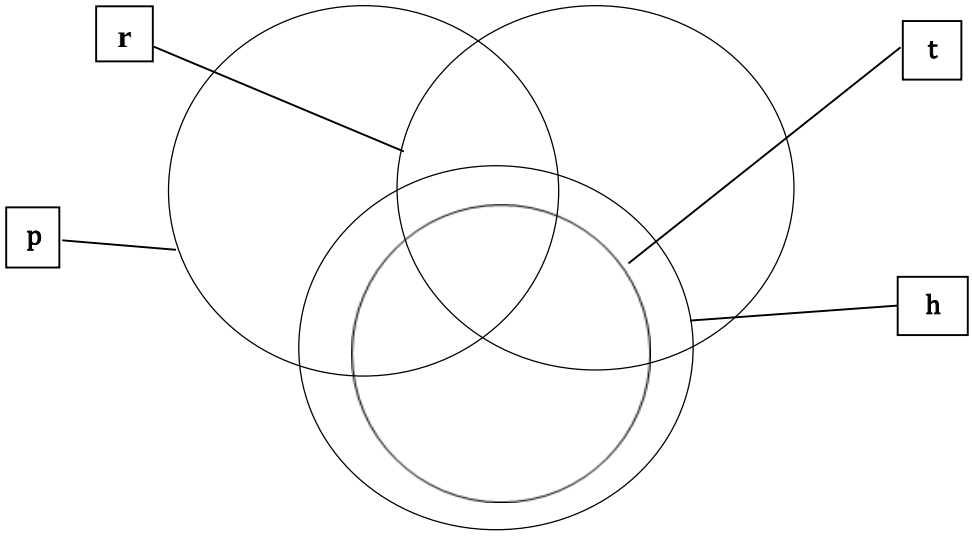


Figure 4.3: Variance Partition Diagram for the Observation Design p×r(t:th)

The above Venn Err variance partition diagram describes facets interrelationships where person facet is crossed with rater facet and where both are crossed with two nested facets of task and themes denoted as **p×r(t:th)**.

The second stage in conducting G studies is the estimation design whereby facet status, random, finite, and infinite, is to be determined for each of the three G study designs described earlier.

4.10.2.1. Estimation Design

At this stage, the first two-facet fully crossed design is composed of three crossed facets, persons-by-tasks-by-raters ***(p×t×r)*** with persons as object of measurement. It has different levels and facets universes which are particularly random and unidentifiable. Persons, raters and tasks are randomly sampled from an indefinitely large universes; they constitute random and indefinite or unidentified facets. That is, persons are selected from millions of possible students; the eight tasks are drawn from millions of admissible tasks; and raters are selected from a sample of indefinable raters that have a potential to score the tasks. The following table (Table 4.17) presents observation and estimation designs embedding facets and their abbreviations (symbols), their observed levels and admissible observations and even their sampling status.

Facets	Abbreviation	Observed levels	Admissible observations	Type
Students	**P**	113	Infinite	Infinite random
Tasks	**T**	2	Infinite	Infinite random
		8	Infinite	Infinite random
Rater	**R**	2	Infinite	Infinite random

Table 4.17: Observation and Estimation Design for P×T×R (With Total Tasks and Subtasks)

The second estimation design includes similar facets as those incorporated in the above observation designs. These facets of student, raters, and themes, as it has already been mentioned, are random indefinite (or infinite random) as they have unlimited number of levels. The facet of theme, however is finite; 04 themes were sampled from the 06 themes assigned in the textbook *'New Prospects'*. The same can be said for topics that can be sampled from infinite universe, but considered somehow fixed because only 04 levels among six (06) are observed in this study as the sampling process is selective

and not randomized. We explored only the four themes exposed to literary/foreign languages streams and discarded those two assigned to teaching scientific streams as almost all ENSB learners were enrolled in literary classes. But this does not mean that scientific streams topics are totally excluded from the study since there are two topics among the four which are shared between the two specialties (those of Ethics in Business and Education in the World), and those very few students stretching from scientific streams also showed their willingness to participate in the study.

In this particular situation, themes can also be considered infinite as they can be among an indefinitely large number of themes existing in the universe. Other themes like pollution, environment, poverty, charity, … etc. can have equal chance to construct the tasks content and situations, and topics can be random indefinite being a corpus in the total vocabulary repertoire, which can be encountered by students in real life communicative situations. Yet, the research purpose endeavors to explore the new bachelors' lexical competence, that is why more focus is put on themes of their textbook; therefore, themes as a facet of measurement is considered finite and fixed in the second and third G study designs.

Facet	Symbol	Observed levels	Admissible levels	Type
Student	**P**	113	Infinite	infinite random
Task	**T**	8	Infinite	infinite random
Theme	**H**	4 (from 06)	Finite	finite fixed

Table 4.18: Observation and estimation designs for p(t:h) design

The third estimation design contains three random facet indefinite universes among the possible universes existing in the reality, and one finite fixed facet, the themes. It considers students, tasks and raters selected among millions of students, tasks and raters that can be potentially regarded in the study. Thus, the sampling status is random infinite for persons, tasks and raters and fixed finite for themes. Information on design three are put forward in Table 4.19.

Facet	Symbol	Observed Levels	Admissible levels	Type
Student	**P**	113	Infinite	Infinite random
Task	**T**	8	Infinite	Infinite random
Rater	**R**	2	Infinite	Infinite random
Theme	**H**	4 (from 06)	Finite	Finite fixed

Table 5.19: Observation and estimation Design for p×r(t:th)

4.10.2.2. Measurement Design

After observation and estimation designs have been explained so far in the aforementioned sections, this section is devoted to a description of the measurement design applied in the current study, the aim of which is to distinguish between students' levels of performance in different vocabulary communicative situations, using a variety of tasks **(t),** raters **(r)**, and themes **(h).** The measurement design is also intended to differentiate between students' scores to make relative and absolute decisions as students are treated as objects of measurement or differentiation facet, and other facets like tasks, raters and themes are considered sources of variance or instrumentation facets.

Corresponding to the measurement designs used in the EduG program, the present study falls upon the following measurement designs:

- The first design: students/tasks raters symbolized by **P/TR**

- The second design: students/tasks themes symbolized by **P/TH**

- The third design: students/tasks raters themes symbolized by **P/TRH**

4.10.2.3. Design Evaluation

This phase of generalizability analysis is interested in estimating the generalizability coefficients that allow the researcher to make decisions about the evaluation process. G theory distinguishes between decisions based on a comparison of students' scores/performance against other students with the aim of depicting students' relative position among other students (relative decisions), and other absolute decisions

based on a comparison of students' scores against a standard or competence of solving problems related to vocabulary knowledge. In this study, therefore, we put emphasis both on relative and absolute measurements because we used relative generalizability coefficients to make relative decisions and absolute generalizability coefficients (dependability coefficients) that fit absolute decisions. Overall, since most of the study facets are infinite and sampled randomly, two coefficients were calculated: a coefficient of relative measurement $\mathbf{Ep^2}$ ("G" coefficient as defined by Cronbach et al., 1972), and a coefficient of absolute measurement. The G coefficient allows the researcher to estimate precisely the degree to which the assessment procedure can place objects of measurement relative to one another. The coefficient of absolute measurement which is defined as "dependability" coefficient, Φ (Brennan, 2001, p. 35) "evaluates the ability of a procedure to locate individuals or objects reliably on a scale in absolute terms" (Cardinet et al., 2010, p. 29).

Relative and absolute generalizability coefficients are various. They depend on the type of observation and measurement designs. This is the reason why equations used to estimate generalizability coefficients implemented in the study are varied according to the number of facets incorporated and the nature of decisions intended from the measurement process and treatment of all facets levels. Eventhough, the *EduG* program was utilized in generalizability analyses, we will introduce relative and absolute generalizability coefficient equations for each study design. We will only present the generalizability coefficients formats and equations without mentioning methods applied in the estimation of variance components because of their variability and complexity.

Relative G coefficient $\boldsymbol{Ep^2}$ **:**

$$\frac{\sigma_p^2}{\sigma_p^2+\frac{\sigma_{pt}^2}{n'_t}+\frac{\sigma_{pr}^2}{n'_r}+\frac{\sigma_{prt,e}^2}{n'_t\,n'_r}}$$

Absolute G coefficient Φ **:**

$$\frac{\sigma_p^2}{\sigma_p^2+\frac{\sigma_t^2}{n'_t}+\frac{\sigma_r^2}{n'_r}+\frac{\sigma_{pt}^2}{n'_t}+\frac{\sigma_{pr}^2}{n'_r}+\frac{\sigma_{tr}^2}{n'_t\,n'_r}+\frac{\sigma_{prt,e}^2}{n'_t\,n'_r}}$$

4.10.2.4. Optimization Design

This stage of G analysis involves defining an optimization design or decision study (D study). It consists of changing the measurement procedure; a set of procedures had been made namely decreasing the facets of the measurement situation, by decreasing the number of tasks, raters, or themes. This procedure aims at identifying the number of facet levels necessary to obtain an optimum level of generalizability and dependability indexes. These modifications made on observation, estimation, and measurement designs might result in an increase or decrease of universe score variance and eliminating or maximizing generalizability variance components (reducing one or more levels of a facet). Accordingly, among these modifications choices, we opted for a procedure that best helped us to achieve more dependable measurements generalizable on different conditions of a measurement situation.

4.10.3. Data Analysis Procedure

After data collection had been accomplished from the research instrument, and after data collection designs had been delineated, the computer program **EduG** was used to estimate not only the variance components for the main and interaction effects for examinees, tasks, raters, and themes but also the generalizability coefficients (relative G coefficient and absolute G coefficient). **EduG** is utilized in our research for data analysis for other two reasons stated by the Swiss Society for Research in Education Working Group (SSREWG, 2006). First, the **EugG** interface is simpler for the casual user to master. Second, EduG is a software package designed to perform generalizability analysis. It is likely, therefore, to prefer **EduG** software rather than **SPSS** or **GENOVA** for an occasional analysis for the aforementioned reasons.

Subsequently, the researcher used **EduG**, the G theory software package, to analyze the quantitative data set obtained from the productive vocabulary knowledge test because it is easy to use and, hence, facilitates generalizability analysis. Practically, "it offers flexibility in the choice of object of study and identification of instrumentation facets. It also provides G coefficients that are appropriate in terms of the particular sampling status of the facets involved" (Cardinet et al., 2010, p. 37). Other available software packages like **SPSS** and **GENOVA** are not favored in this study. Although

SPSS can estimate the variance components, it cannot estimate generalizability parameters. This is on one hand; on the other hand, **GENOVA** is difficult to handle and not flexible to achieve generalizability analyses.

The Swiss Society for Research in Education Working Group (SSREWG, 2010) describe **EduG** as that software designed to conduct generalizability analysis which is bound to **ANOVA** (Analysis of Variance) and G theory. EduG is also defined as a windows-based program developed for generalizability analysis purposes (Cardinet et al., 2010; SSREWG, 2010), created by Cardinet and his associate Bertrand. It is available in two versions: French and English but the one used in this study is of English version. It is possible to download and install it for free from this website http://www.irdp.ch/edumetrie/logiciels.html

EduG determines the sources of variance that have strong effects on observations. It also determines the effect of changing the observation design intended to eliminate principal contributions to measurement errors variance. Accordingly, it permits the researcher to estimate the reliability of the measuring instrument in paly. **EduG**, "helps you to see how to change your design to achieve a higher degree of reliability in future measurements" (SSREWG, 2010, p. 37). **EduG** is a statistical program used for calculating and presenting results gained from generalizability analysis, but it is the researcher's role to interpret those results. However, this tool requires careful knowledge and understanding of analysis of variance and G theory for more professional usage in order to yield accurate results resembling to those of **GENOVA**, developed by Crick and Brennan (1983). **EduG** developers assume that it is an alternative to **GENOVA** as it is flexible and easy to use and have attractive interface (Clausser, 2008; SSREWG, 2010).

Following the same logic, Clausser (2008) suggest that **EduG** helps in carrying out generalizability analysis by inserting raw data (here observed test scores) or the sums of squares obtained from analysis of variance. Even so, the results obtained from **GENOVA** are still useful up to date, but the development of information technology made it difficult for **GENOVA** to make a psychometric study when compared to **EduG.** Furthermore, the latter is said to be "unique among other available generalizability

software packages in that it was conceived specifically to exploit the symmetry property of G theory" (Cardinet et al., 2010, p. 37). The symmetry property that was first introduced by Carindet, Tourneur and Allal in 1976 is characterized by its flexibility, in that it allows the choice of object of study, and identification of instrumentation facets. Besides, Cardinet et al. state that the symmetry property presents G coefficients that are suitable for the particular sampling status of the facets involved in the study design.

EduG is a Windows-based or Vista-based software, easily useable, free downable, and "readily obtainable" (**SSREWG, 2010, p. 37**). Another characteristic feature of the program is its ability to estimate variance components before application of **Whimbey**'s correction factor of the sampled facets from non-infinite (fixed) universes (see Chapter one three for a full description). Using **EduG**, it is possible to conduct generalizability analyses, basically, either by scrutinizing the EduG user guide published by Swiss Society for Research in Education Working Group (SSREWG, 2010), or consulting an important reference entitled *"Applying Generalizability Theory Using EduG"* authored by Cardinet, Johnson, and Pini in 2010. It is so easy for the user to move between the program interfaces and enter the data set and obtain the results, yet, what is left how to interpret them correctly.

Launching a G study requires a thorough insertion of the observed data or sums of squares in the program. After that a set of procedures can be requested: the calculation of level means, analysis of variance and G study analysis, estimating **Phi** coefficient **(Lambda)**, a D study optimization analysis, and a series of analyses relevant to modifications designs (Clausser, 2008; Cardinet et al., 2010; SSREWG, 2010).

The EduG produces conventional tables that present values of the means to be obtained for all the possible distinct observations of sources of variance; for each level of each facet such as students, tasks, raters, and themes, and for each facet interaction, like tasks by raters, students by tasks, students by themes, students by rater. These statistical values can be calculated by EduG and displayed in analysis of variance and generalizability analyses tables. The ANOVA table summarizes usual sums of squares, degrees of freedom, and mean squares, variance components, estimated standard deviation for each facet level, and the proportion of the variance of individual scores

that is attributable to each facet. In the G study table, the second table of G analysis, the contributions to each type of variance to the students' true score, relative error variance, absolute error variance, total true score variance, absolute error and relative error variances and their standard deviations, G coefficients and dependability coefficients are presented.

Decision studies described as optimization studies tend to improve the quality of the current measurement tool, the test, or to increase measurement reliability and reduce the measurement error by means of optimization and G-facets analysis. It estimates the G study results of the conditions in which facet levels sound to be distinct from observed data. Data thus observed allow, again, for an estimation of score generalizability when a given level of a given facet is neglected or eliminated. Optimization table, therefore, responds to the question, for example, what if we eliminate or increase the third rater, and or what if we reduce the fourth theme, or the third task. In addition to these options, in G-facet analysis EduG can analyze the scores or data related to female and male separately as it can re-make data analysis by changing facet characteristics in the estimation design process. Since students are considered object of measurement (differentiation facet), we do not consider those significant statistical differences in lexical performance as we focus on the relative effects of tasks, raters, and themes on scores reliability. To illustrate, designs with no differentiation facet considers fixed facets instead of random facets (fixing facets).

The present study seeks to apply the principles of G theory; the different stages used in the assessment of the measurement situation, and the EduG software henceforth, perusing the same logic as set up forward in the literature by Cardinet, Johnson and Pini in 2010. The EduG program is characterized by its straightforwardness in importing (computing) data and its flexibility and easiness in the use of its work screen. Having said, applying G theory using EduG is not an easy task as it requires a thorough understanding of and control over the principal terminology of the theory; otherwise the study will never be fulfilled. This difficulty lies in the procedures used to determine the different study designs in a correct way, that in turn determine G analyses. Care,

therefore, should be put on observation design, estimation design and thus for measurement designs that form the basis for any G study.

In the first phase, termed observation design, it is required from the part of the researcher to determine the crossing relationship existing between facets, their interactions or inter-relationships, and their levels in the data set. Next, in the estimation design phase it is necessary to reconsider facets sampling status because facets can be treated as “fixed”, “random,” or “finite random” in a measurement situation as it is also important to determine their number of levels in the universe of admissible observations. The more correct these designs are the more accurate information, we would have in carrying out G analyses, particularly, partitioning total variance appropriately in analysis of variance, and estimating variance components correctly. Third, the measurement design determines the facets representing the objects of measurement ‘differentiation face’ and ‘instrumentation face’ that embody a measurement procedure (for more knowledge on the EduG work screens used in this study see Appndix G).

The analysis of the quantitative data, therefore, aimed to assess the present assessment; to investigate the assessment procedure in terms of accuracy (validity) and consistency (reliability) by means of estimating the variance components of tasks, raters, and themes that might affect the measuring instrument and might contribute to measurement error to draw a complete picture of the assessment precision. The data being analyzed would consequently help in making decisions when optimizing measurement precision and dependability from G theory perspectives and to better understand G theory principles.

Conclusion

This chapter included an account of the practical and methodological procedures used for the description and identification of the descriptive method that fits into the nature of psychometric studies that aim at discovering the variance components affecting the generalizability and dependability of lexical competency test scores. It described the study sample and data gathering instrument. It also described out the different procedures applied in the design of the performance test after the intended

competency (vocabulary knowledge) had been delineated. The scoring guide and rater training procedures were also explored in this chapter and holistic rubrics were used to quantify observation. As to data analysis procedures, following G theory principles, this chapter described the data collection study designs and the EduG software program applied to analyze data quantitatively. Generalizability principles contained typically five design types: observation design, estimation design, measurement design, design evaluation, and finally optimization design. The coming chapter reports the findings obtained from the actual research measuring instrument used to obtain and quantify students' performance.

CHAPTER FIVE

ANALYSIS AND PRESENTATION OF THE RESUTS

Introduction

This chapter analyses and displays the research results in an attempt to answer the six major research questions posed, with their corresponding facets and levels of measurement, designed in the G and D studies. It provides the descriptive statistics relevant to the facets of the measurement situation by estimating the means and the standard deviations of the vocabulary observed test scores. This chapter also aims to quantify, evaluate and even improve the psychometric properties of the observed test scores, namely reliability and validity, obtained from the performance test of depth of productive vocabulary knowledge. The quantitative data, thus, gathered and their analysis exhibit an assessment of the overall reliability of scores, and of the study method henceforth. To this end, percentages of test scores variation, sources of measurement variability contributing to measurement error, generalizability and dependability of test scores for relative and absolute decision making, estimation of validity, reliability and dependability for alternations made to measurement procedures with regard to facet levels sampled from the universe of admissible observations to attain optimal degrees of reliability and score consistency are further computed and explored. Driven by these study procedures, the research findings and data sets are to be presented, interpreted and justified in a way relevant to previous research.

5.1. Data Analysis and Presentation Procedures

We conducted separate, but related, G and D studies using data gathered at two time points and two types of questions have been raised: three are associated with the G studies and the other two questions are linked to the D studies[7]. Subsequently, data analysis is organized around the research questions and their corresponding study designs. Each question has a significant generalizability design estimating sources of variability and G coefficients and the same does hold true for our D studies, where optimization procedures have been made to reach an optimal reliability index to achieve a certain magnitude of the G coefficients for absolute decisions.

The first phase of data analysis is concerned with three G studies and the second phase deals with multiple D studies. Initially, the G studies entail one fully-crossed, and two partially nested or crossed designs. The fully-crossed two facet random person×task×rater generalizability study design model aims to answer the first question related to investigating the relative impact of persons, tasks and raters on the test scores and learners'overall lexical performance. As such, this design identifies the estimated variance components and their percentile effects on the students' universe score. These sources of variance include: person (P) variance, task (T) variance, rater (R) variance, student-task (PT) interaction variance, students-rater (PR) variance, task-rater (TR) variance, and the residual component (person-task-rater (PTR, e), which denotes uncontrolled indefinable or unmeasured sources of error).

The second partially-nested two faceted design integrates the facet of theme (H) into the whole G study design, the aim of which is to estimate sources of variability and their effects on the students' universe scores, mainly theme main and interaction effects on the overall vocabulary score reliability. Within this measurement model, the estimated sources of measurement error involve: theme main effect (H), tasks nested within themes (T:H) effect, person crossed with themes (PH), person-by-task interaction nested within themes, and a residual effect (PT:H, e). It is worth mentioning, here, that it is not possible to estimate all the variance components separately because of the nesting relationship and facet interaction effects. Student or person main effect is not considered because this facet is the object of the present measurement, it is not treated as a source of error, it is rather a facet of differentiation, and the same can be said for all the three G study designs.

Note that although the universe of admissible observations is crossed in nature, the researcher has chosen a nested rather than a crossed design. This decision has been made based on Shavelson's and Webb (1991) insight, which confirms that occasionally facets are nested in G studies because of cost or logistic considerations. In the present study, the choice is dictated by logistics, because the tasks are categorized under the theme headings, tasks are treated as sub-sections of themes, they are nested within themes; eventhough every task was administered to every student and every rater scored

every student's task performance, and thus for themes. One might assume that, this facet should keep pace with a crossed measurement but since we seek to identify the variance components, including themes as a facet and its effects on the overall test performance and universe score, we opted for a nesting design to examine the effects of themes on task performance and on rater's judgements. The second and third designs are respectively partially-nested and partially-crossed because they do have both crossed and nesting effects on students' performance.

With the third measurement model, variance components can be estimated for persons, raters, themes, tasks nested within themes, persons-by-raters, persons-by-themes, persons-by-task effect nested within themes, rater-by-theme effect, rater-by-task effect nested within themes, person-by-rater-by-theme interaction effect, person-by-rater-by-theme interaction effect nested within themes and a residual, undefinable unmeasured sources of error. The variance components abbreviated as (P, R, H, T:H, PR, PH, PT:H, RH, RT:H, PRH, PRT:H, e respectively) are identified in our G studies to estimate as much as possible sources of measurement variability and their relative effects on the universe-score variance. Compared to a completely crossed design, various variance components can be estimated within this crossed-nested design.

It is possible to sum up how data analysis is approached in the G studies phase along the following table (Table 5.1), which involves the G studies and their correspond

Facets	Levels	Research Questions	Design
Students	**113**	What is the relative effect of persons, tasks and raters on the generalizability and dependability of test scores obtained from the vocabulary performance test?	**p×t×r**
Tasks	**08**		
Raters	**02**		
Students	**113**	What is the relative effect of persons, tasks and themes on the generalizability and dependability of test scores obtained from the vocabulary performance test?	**P (t:h) or P(h:t)**
Tasks	**8**		
Themes	**4**		
Students	**113**	What is the relative effect of raters, tasks and themes on the generalizability and dependability of test scores obtained from the vocabulary performance test?	**Pr (t:h)**
Raters	**2**		
Tasks	**8**		
Themes	**4**		

Table 5.1: G Study Designs, their related Research Questions and Facet levels

-ing facets and the number of levels sampled from the universe of admissible observations, the research questions and the corresponding study designs.

The second stage termed optimization design (D studies), however, is not concerned with estimating sources of error variance, it rather aims to make changes pertinent to the *"what if analysis?"* procedure (Cardinet et al., 2010) to arrive at a certain level of score consistency and dependability of the research measurement procedure. It definitely starts with the results obtained from the G studies especially when the G coefficients are acceptable or increasingly high and good. In this particular stage, more emphasis is put on the optimization designs and the type of decisions made by the researcher as concerns facet level sampling alternations or reductions involving the facets of tasks and raters. Succinctly, this section is an attempt to answer the question what if we reduce the levels of raters and the levels of tasks (see Table 5.2 underneath). Overall, our D studies endeavor to answer the research questions related to decreasing the number of tasks and raters and their effects to attain optimal G coefficients. Data analysis of the D studies are based on the results obtained from the various optimization designs. The current D studies relevant facets sampled levels and the corresponding research questions and designs are summarized in Table 5.2. The Optimization procedure serves to reduce the cost effect of data collection especially by reducing the number of tasks and judges.

Facets	Levels	Research Questions	Design
Students	**113**	What is the effect of decreasing the number of tasks designed to assess vocabulary performance on the generalizability and dependability of test scores?	**p×t×r**
Tasks	**From 08 to 1**		
Raters	**02**		
Students	**113**	What is the effect of decreasing the number of raters on the generalizability and dependability of test scores?	**p×t×r**
Tasks	**From 08 to 01**		
Raters	**From 02 to 01**		

Table 5.2: D Study Designs and their Corresponding Research Questions, Facets and Numbers of Levels

The two questions posed in the Table above are investigated within the D studies that aim to:

1. Investigate the extent to which the vocabulary test is dependable in terms of G coefficients for relative and absolute decisions.

2. Investigate how many items themes are optimal to achieve a certain magnitude of G coefficients for relative and absolute decisions.

Setting up three G studies together with four D studies we will strive to answer the basic six research questions together with the sub-questions. To conduct data analysis procedures, the EduG software program was used as a G calculator to compute the percentages, the mean scores, the standard deviations, the variances, the sums of squares for each dependent variable using ANOVA (Analysis of Variance) with tasks, raters, and themes as varied factors, the G coefficients, and the dependability coefficients. This computer package also enables the researcher to estimate the variance components along with the standard errors of measurement (SEM). In the EduG software, we indicated person (P) for students as an object of measurement, with tasks (T), raters (R), and themes (H) as the measurement facets. Students, tasks, raters and themes were randomly sampled from the universe of admissible observations; from a population of students belonging to the Department of English at ENSB, and eight tasks were sampled from the universe of admissible tasks, and thus for raters. As to the themes, it is indicated that they were finite random sampled from the possible themes that can be used to contextualize the tasks at hand. For the D study procedures, the EduG software estimated the coefficients, relative and absolute, for the numbers of tasks and raters to find out the number of tasks and raters necessary to achieve an acceptable degree of score reliability.

5.2. Analysis and Presentation of G Study Results

The results obtained in the G studies are presented through three phases; each displays one separate study design, beginning with the PTR fully crossed design, pursued by the second P(T:H) partially nested design then followed by design three P×R(T:H) which is partially crossed.

5.2.1. Analysis and Presentation of the First Design PTR Results

The results yielded by the PTR design are displayed via four stages: Setting up a G study, where study descriptive statistics for both the raters and tasks facets are provided accompanied by observation and estimation designs stage, pursued by analysis of variance, the generalizability analysis, and measurement design stages. These appear below.

5.2.1.1. Setting Up the G Study for the PTR Design

The fundamental aim of the current G study is to evaluate the psychometric conditions, specifically reliability and validity, of the measuring procedure through an estimation of measurement precision. Accordingly, the different variance components in play will be identified in the observation and estimation designs phase, and then quantified henceforward in the measurement design phase and finally analysed and interpreted with view to their relative size contribution to the total measurement error and their effects on the measurement precision using ANOVA procedure. The G study, therefore, goes through the three basic phases of observation and estimation designs, ANOVA and estimated variances processes, ended up by the measurement design.

5.2.1.2. Observation and Estimation Designs

In this section the relevant G study observation and estimation designs are displayed for the two-facet fully-crossed random **p×t×r** design, where the subsidiary facets of the study are involved in the application as shown in Table 5.3.

Facet	Label	Levels	Univ.
Person	P	113	INF
Task	T	8	INF
Rater	R	2	INF

Table 5.3: Observation and Estimation Designs for the p×t×r

Table 5.3 describes the study main facets: the facet of students or persons (P) is treated as the object of the measurement when applying the principles of symmetry suggested by G theorists. The number 113 is the level standing for students; the task

facet contains eight levels and raters have two levels representing the facets of both the observation and measurement designs. These dependent and independent variables, meaning facets of measurement are infinite (INF); they are randomly sampled from the universe of admissible observations (Univ). The two facets, namely students and tasks, are each treated as infinite random being selected from an indefinitely (infinite) large population of students and from an extremely large number of tasks and, hence for raters. That is anyone student from the entire population can have the opportunity to be a participant in the current study, and the same does hold true for the eight tasks that can potentially be sampled from within their congruent infinite universes (estimation design). Since all the eight tasks were responded to by every one of the 113 students, both students and tasks are crossed. These will be further presented with their within-subject ANOVA. Raters are also crossed with subjects and tasks because they scored all the tasks completed by all the students.

Note that the EduG program originally provides users with the fifth column where they can opt for facet or facet level sampling reduction after a clear identification of G study facets and their level, and sampling status within the universe. The column is deleted from the table above because we are not concerned with studying the results of G study after reducing the number of levels. Rather the principal purpose of the G study is to estimate the variance components for the different study facets and effects using ANOVA procedures and establish whether the test of depth of productive vocabulary knowledge could reliably differentiate among first year university learners at ENSB.

The EduG enabled us to obtain values of the means for all the plausible distinct variance components: for every level of any facet and for level interactions across facets. For the case of our measurement situation, the mean is provided for each task and for each rater being instrumentation facets. These instrumentation facets represent the sources of measurement variability accounting for the amount of error contributed to the true score variance. The observed, but not the expected, variance is further provided alongside each mean; the observed variance for values for those components that have contributed to the mean. The same does hold true for the variances and the standard deviations.

5.2.1.3. Descriptive Statistics for the Study Facets

In an attempt to answer the research questions, we were in a position to provide descriptive statistics for the major facets formulating the basis of the current measurement. These statistical descriptions permited a deeper understanding of the study designs and facet levels sampled from the universe of admissible observations. We begin with exposing the descriptive statistics related to the facets levels of the measurement situation associated with estimating reliability and then, the same procedure goes for estimating validity. To do this, the mean scores, the standard deviations of the tasks and raters (and themes in the second and third designs) facets and their respective levels are presented. The EduG files and tables containing the data, are provided by the software package for each observation, estimation and measurement design as sub-sections related to the three research designs. The design information provided by EduG in the work screen are reproduced in the form of statistical tables.

5.2.1.3.1. Descriptive Statistics for Tasks

This section displays the descriptive statistics set for the overall tasks included in the first two-facet fully-crossed random research design. It describes the tasks variable performed by the participants. The results, thus, obtained are presented in Table 5.4 below, where the mean scores, the variance scores and the standard deviation for each task of the eight tasks are handed out.

Task	Mean	Variance	Std. Dev.
1	2.084	1.068	1.034
2	1.925	0.804	0.897
3	1.695	0.778	0.882
4	1.664	0.905	0.951
5	1.832	0.910	0.954
6	1.743	0.766	0.875
7	1.752	0.779	0.883
8	1.646	0.919	0.959

Table 5.4: Descriptive Statistics for Tasks Facet in the ptr Design

Note. Grand mean for levels used: 1.793; variance error of the mean for levels used: 0.886; standard error of the grand mean: 0.941.

Descriptive statistics in Table 5.4 display the values related to the persons' mean scores across the eight tasks (the expected mean of the level of performance). The mean, the task performance mean scores, of the current research dataset is the sum of the overall scores divided by the number of students.

Holistically, the test takers performed (scored) distinctly through the eight tasks. Compared to the grand mean for levels (1.79), the participants showed very weak performance in some tasks (Task 08 by 1.64; Task 04 by 1.66; Task 03 by 1.69) and relatively low in some others (Tasks 07 and 06 by 1.74 and 1.75 respectively). The average scores, however, for tasks 05 (1.83) and Task 02 (1.92) are higher than the expected mean scores, and proved to be extremely high in Task 01. This arguably explains that the respondents performed best in the latter tasks compared to the previous ones especially in Task 01. The higher the score, the more simple the task would be and vice versa. In addition, the students' average scores are variant across the tasks; the standard deviations are different and this, indeed, resulted in a considerable variance in the tasks and student-task interaction variance componentds.

5.2.1.3. 2. Descriptive Statistics for Raters

In this section, descriptive statistics is used to account for the means and standards deviations per each rater across the eight tasks being the source of lexical performance judgments.

R (Rater)	Mean	Variance	Std. Dev.
1	1.826	0.922	0.960
2	1.759	0.847	0.920

Table 5.5: Descriptive Statistics for Raters in the Fully-Crossed ptr Design

Note. Grand mean for levels: 1.793; variance error of the mean for levels: 0.886; standard error of the grand mean: 0.941.

As indicated above in Table 5.5, the mean judgments for the first rater is relatively

high (1.82) in comparison with the second rater (1.75). Even so the mean weightings are not so far distant but approximate between the two raters which means that they may have scored the tasks consistently; thus the consistency of the observed test scores across raters is achieved. Besides, it might be observed from this table that the standard deviations are closer alongside the raters' judgments.

Although the tasks have arguably proved to be at acceptable level of difficulty and that the raters were consistent in weighting the tasks, this includes a relative measurement for both task and rater facets. It needs to be consolidated by exploring the reliability of an absolute measurement in a G study.

5.2.1.4. Analysis of Variance: A Generalizability Analysis

After the facets of the current measurement situation have been identified together with their interrelationships and sampling status, an analysis of variance is required to obtain more information about study facets. ANOVA is conducted to estimate the variance components at paly in order to elicit the current measurement reliability and estimation precision. G analysis begins right at this point, and the results are presented as under in Table 5.6.

Source of variance	SS	Df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	903.158	112	8.064	0.451	0.451	0.451	50.5	0.067
T	35.190	7	5.027	0.015	0.015	0.015	1.7	0.011
R	2.058	1	2.058	0.001	0.001	0.001	0.1	0.002
PT	467.373	784	0.596	0.211	0.211	0.211	23.6	0.016
PR	47.879	112	0.427	0.032	0.032	0.032	3.5	0.007
TR	8.893	7	1.270	0.010	0.010	0.010	1.1	0.005
PTR, e	136.669	784	0.174	0.174	0.174	0.174	19.5	0.009
Total	1601.221	1807					100 %	

Table 5.6: Analysis of Variance for the p×t×r Design

Note. SS: sum of squares, Df: degrees of freedom, MS: mean squares

Table 5.6 presents the proportion of variance of individual scores. The present focus, however, is put on checking the impact of the different sources of error variance on the students' mean scores but not on individual scores. To this end, we opted for a measurement design where students (p) are considered a differentiation facet and tasks and raters are instrumentation facets. Since the actual measurement situation is not in favour of analysing the relative effects of variance components on individual scores, Table5.6 above, even important to mention, will not be analysed statistically in details, it will rather be explained in global terms. Cronbach (1951), in this regard, recommended that these percentages should not be interpreted "as directly reflecting the relative importance of each variance source, since real life decisions are generally made on the basis of total scores or mean scores, and not on the basis of non-summarized data points". The relative magnitude of each source of error in the total error variance is rather given in the upcoming G study table (Table 5.7, p. 248) partitioning the relative error variance and absolute error variance across the different facets and their interaction effects.

The ANOVA table, Table 5.6, is the conventional table in G studies. It defines the universe sizes and provides estimates of the observation and estimation designs for the (p×t×r) fully-crossed design. For our case estimates of person, task, rater facets and their corresponding interaction effects, pt, pr, tr, ptr (read as person-by task interaction, person-by-rater interaction, task-by rater interaction, and person-by-task-by-rater interaction) are thus far illustrated in the Table.

The ANOVA table, by far, provides us with the data set related to the usual sums of squares, degrees of freedom, and mean squares for each facet and facet interaction. Furthermore, it gives sets of information about the estimated variance components. As they appear in Table 33 above, each data set is related either to a random effects model illustrated in *Random* column, or to a mixed model indicated in *Mixed* column, and or to a third set of corrected components in the column entitled *Corrected,* which shows variance component values for the p×t×r design after Whimbey's correction factor have been applied (see Chapter 3).

The (SS) column represents sums of squares. In statistics, sums of squares assign variability or variation to distinct variables in a sequential order, that is why it is also known as sequential sums of squares. Table 5.6 (p.44)shows that a maximum variation refers to student (P) by 903.158 which deviates from the mean value. A higher sum of squares explains higher variability. This variation is followed by the interaction effect variance of PT by 476.37; to the interaction effect of PTR (136.66). The remaining sums of squares range from 47.87 for PR; 35.19 for tasks; 8.89 for TR and 2.05 for R indicate low variance compared to mean value and the other facets and facet interaction effects. The rest variation is assigned to the residual (unmeasured sources) sums of squares.

In the second column headed df, degrees of freedom have also been computed as degrees of (n-1) for both instrumentation facets (df= number of tasks -1 or df = 8-1 and df= the number of raters –1 or df=2-1) and differentiation facets (df= number of persons -1 or df =113-1). Degrees of freedom are calculated for the sake of computing the mean squares (MS) listed in the third column entitled MS. Any mean square value is calculated by dividing the sum of squares to the degree of freedom in order to compute the variance components.

The penultimate or the percentage "%" column 7 displays the proportion of the variance of individual scores, here, estimated as the sum of the corrected components of the random model. It is worthy to mention that the relatively high contribution to the total score variance of the facet students, along with the interaction effect between students and tasks and raters is confounded with unmeasured sources of variance. The major variance contributions are illustrated as under in Table 5.6 (p.244). According to SSREWG (2010), the data set provided by the software should be considered as indicative only, eventhough they may be invested to establish confidence intervals to test the significance of the variance components. It also confirms that the standard errors in paly are associated with the random effects model components (those of column 5, rather than those of column 7).

Overall, the ANOVA results are denoted in Table 5.6. By far the greatest contribution to the total score variance, at just over 50 %, comes from students (P), then from students interaction with tasks (PT) by 23.6%. PT variation is, therefore, the next

largest variance contributor, followed by students interaction with tasks and raters confounded with the residual (P-by-T-by-R, E) by within 19.5%.

5.2.1.5. Measurement Design

The G study Table (Table 5.7, p. 285) sums up the measurement design for the p×t×r results, where students (P) are considered as a differentiation facet, tasks (T) and raters (R) are instrumentation facets. The results represent the effect of sources of error on the students' vocabulary mean scores. The table disentangles the relative error variance and absolute error variance along the facets and their interactions. To estimate the SEM, the differentiation facet must be determined of infinite size, as G theorists and practitioners often confirm, and it is represented by the facet of students (P) in the present measurement situation.

Before conducting a generalizability analysis, it is salient to describe the two-crossed facet person × task × rater design generalizability analysis table (Table 5.7) to the reader to facilitate internalizing data analyses procedures. In the subsequent designs namely Design 2 and 3, we will consider only the values reported in the tables.

To begin with, the first column headed differentiation variance accompanied by the second column entitled source of variance together display how the sources of variance are disentangled in the actual research measurement design. They either contribute to the true (i.e. differentiation) score variance or to the relative or absolute error variance. In each column contributions of each error variance to measurement are provided. The percentage column for "% relative" or "% absolute" indicate the relative influence of every source of variance for relative and absolute scale measurements. The raw column, horizontally provides estimates of the sum of variances both for the true variance and the error variances for relative and for absolute measurement.

The penultimate column sums up the standard deviations, being equated to the standard error of measurement, they can be used to determine a confidence interval around the true mean score for each object of measurement. For Cronbach this sort of information is the most important since it can be directly interpreted, for both relative and absolute measurement.

Source of variance	Differentiation Variance	Source of variance	Relative error variance	% Relative	Absolute error variance	% Absolute
P	0.451		.....		.....	
	.....	T	.....		0.002	3.3
	.....	R	.....		0.000	0.5
	.....	PT	0.026	49.7	0.026	47.2
	.....	PR	0.016	29.8	0.016	28.3
	.....	TR	.....		0.001	1.1
	.....	PTR,e	0.011	20.5	0.011	19.5
Sum of Variances	0.451		0.053	100%	0.056	100%
Standard Deviation	0.672		Relative SE: 0.230		Absolute SE: 0.236	
Coef_G relative	0.89					
Coef_G absolute	0.89					

Table 5.7: Generalizability Analysis for the *ptr* Design

Note. Grand mean for levels: 1.793; Variance error of the mean for levels: 0.007; Standard error of the grand mean: 0.085. Dots (a row of dots) appear in the above table represent null values indicating an observation design reduction.

In the lowest section of the G study table are the final results, representing the two generalizability coefficients, relative (Coef_G relative) and absolute (Coef_G absolute). The relative coefficient (Coef_G relative) considers the sources of variance affecting a relative scale of measurement and the other coefficient accounts for sources of error related to absolute measurement scale. These coefficients sum up the data of the two tables (Table 5.6 and Table 5.7) of ANOVA and of G study upon which the EduG user can easily interpret the quality of the overall measurement design on a reliability scale ranging from 0 to 1.

The result in the current case is SE: 0.230 for the relative SEM and SE: 0.236 for the absolute SEM. Differences between these values, as can be noticed in Table 5.7 are negligible, despite this they are important to be aware of. Because, measurement error

is a usual form of inaccuracy, it can differentiate between a measured score and its true score.

Table 5.7 further indicates that the largest error variance component is for person-by-task interaction (PT), for both relative measurement (by 49.7%) and absolute measurement (by 47.2%), and the error variances show that the mean score of individual performance of some students is high in certain tasks and relatively low in other tasks. This can be explained by task variation with different degrees of facility and difficulty, except for task 01 that proved to be easy compared to others (see Table 5.4), and the reverse can be said for some other students. The second source of error variance depicted in students-by-rater interaction (PR) by 29.8 % for relative measurement and 28.3% for absolute measurement, which demonstrates the ratings variation across individual performance; once the student changes the raters' judgment also changes. Differences in assessing performance from one exam sheet to the other may refer to the student's language to do the task which may have a positive or negative influence on the scorer' attitudes and emotions. The third variance component having an increased effect on the actual measurement precision is the PTR,e (students-by-task-by-rater interaction confounded with all unidentified sources of variance (e)). Its proportion of the total variance is 20.5% for relative measurement and 19.5% for absolute measurement. In the meanwhile, the remaining sources of variance proved to be negligible, as having less impact on score differentiation.

Compared to their effects on the score variability, the between-rater variance (0.5%) is very low which means that the two raters were consistent in their weighting of the students' performance on the quality of word knowledge. As such, inter-rater consistency has been arguably proved in the application of the scoring rubrics and in providing appropriate judgements on individual performance or students responses, followed by TR (task-by-rater interaction) by 1.1%, which explicates very low variation in scoring the task performance; ratings are not relatively very distant alongside the tasks. The between task variance proportion (or the main effect for tasks) accounts for 3.3% indicating that all the eight tasks are different but approximate, or rather of equal level, in their degrees of facility and difficulty except for Task 1 that proved to be facile.

This might be explained by the fact the tasks were almost accessible to all learners because they are acquainted with knowledge about the themes of the textbook units and are familiar with the target words being exposed to during a whole course of instruction even put in authentic novel situations.

Note that the results of both relative and absolute error variances are slightly different across sources of variance, and this in turn, evidently interprets no variance in the G coefficients obtained from the G studies. The generalizability coefficient, as seen in chapter one, refers, in brief, to the ratio of the universe score variance and to both the universe score variance and the relative error variance. The G coefficients for person measurement (differentiating individuals not differentiating items) have acceptably high values for both relative (0.89) and absolute (0.89) measurements. Both values extend a coefficient of .80 that has been taken as a criterion value to consider scores reliable (Cardinet et al., 2010). The G coefficients explain how students' scores are dependable and generalizable (satisfactory measurement precision) across test items (tasks) in the overall assessment of vocabulary knowledge whether the results used to compare students with each other or to compare their performance against the expected level of performance.

5.2.2. Analysis and Presentation of the P(T:H) Design Results

The second G study implements a partially-nested P(H:T) design the purpose of which is to examine whether the measuring procedure can differentiate among themes from the perspective of their relative difficulty, or item facility. It also aims to explore the importance of themes as a hidden source of error variance in the dependability and generalizability estimates of the vocabulary assessment scores. The details of the data structure obtained in the second G study are presented in Table 5.8, p. 251.

In the assessment of productive depth of vocabulary knowledge, we found it important to embrace themes (H) as a facet in the measurement situation. The relevant design was a mixed design; crossed and nested, random and fixed. In this two facet mixed P(T:H) design, the H is fixed and the themes per each task differ from theme to theme, so the tasks are nested within the themes. Since all the 113 students responded to all the tasks in all the themes, the design is partially nested, with tasks (T) nested

within themes (H) and evidently both crossed with persons (p). The notation for this design is p×(T:H) or P(H:T). It is worth mentioning that the themes observed in the G study selected purposefully with no intention to generalize beyond them. In this sense, the theme facet is fittingly treated as being fixed.

5.2.2.1. Observation and Estimation Design

The observation design comprises a set of information described in Table 5.8 appearing below, which involves facet identification and numbers of levels observed.

Facet	Labl	Levels	Univ.
Persons	P	113	INF
Themes	H	4	6
Tasks-within- Themes	T:H	2	INF

Table 5.8: Observation and Estimation Designs for the p(t: h) Study Design

There are three facets comprising the second G study design: Persons (P), Themes (H), and Tasks (T) that are declared for item difficulty study. Eventhough every student of the 113 sample had been administered the test eight tasks, and hence the themes underpinning the tasks, all the three facets are partially nested. As stated earlier, although the three facets are expected to be fully crossed, the logistics justifies the nesting interrelationships and at the same time the crossing interrelationship is also approved. Because the eight tasks are categorized under four theme headings, the tasks are nested within the four themes; each two tasks in the G study subcategorized within one theme. Task1 and Task 2 underpin the theme of Ancient civilizations, Tasks 3 and 4 consider the theme of Education in the World, Tasks 5 and 6 tackle Ethics in Business, and Tasks 7 and 8 deal with Feelings and Emotions. The symbol facet T:H means that each two tasks from the eight tasks are nested within one theme.

As to the estimation design, the students (P) facet is considered infinite random as with the first G study design, the facet of theme, however is finite; 04 themes were sampled from the 06 themes existing in the textbook *'New Prospects'*. The tasks and persons as facets of measurement theoretically have no limited number of levels.

Similarly, themes that can be selected from infinite universe, but are considered somehow fixed because only 04 levels are observed in the present study as the sampling process is selective rather than random. The two themes assigned to teaching scientific streams were excluded and only what concerns foreign languages stream are explored; but in this particular case themes can also be considered infinite as they can be among an indefinitely large number of themes existing in the universe. Other themes like pollution, environment, poverty, charity, … etc can have equal chance to construct the tasks content and situations.

5.2.2.2. Descriptive Statistics for Themes

After the measurement conditions have been identified in the observation and estimation designs phase, this section is devoted to calculating the theme mean scores that are statistically described in Table 5.9 below.

Themes (H)	Mean	Variance	Std. Dev.
1	2.097	1.017	1.009
2	1.717	0.858	0.926
3	1.748	0.826	0.909
4	1.739	0.874	0.935

Table 5.9: Descriptive Statistics for Themes in the Two Facet Partially-Nested p(h:t) Design

Note. The grand mean: 1.825, Variance: 0.919, Standard dev.: 0.958.

Table 5.9 clearly indicates that the theme mean scores, the students' performance across themes, are low compared to the grand mean, except for H 1 that of Ancient Civilization which exceeded the expected mean score accounting for 2.097 of the total mean score of 1.825. This arguably confirmed H 1 (task 1 and 2) to be easy compared to the other themes in the remaining six tasks (task 3 through 8) where students' performance varied and sounds weak, ranging between high mean score for H 2 (by 1.717), H4 (1.739) and 1.748 for H 3. This further indicates differences in theme mean scores especially H 4 and H1. Hence, H 1 is easier than H 3, and H 3 is easier than H 4

and H 4 is easy than H 2; the later theme, *Education in the World,* proved to be the most difficult for learners. The measurement procedure could differentiate between themes in terms of their facility. Accordingly, themes can be ordered from easy to difficult starting with H 1, then H 3, then H 4, and then H 2.

The variance column indicates, with respect to the mean scores, the amount of variation existing among students' performance across the four themes. As to the standard deviation, the square root of the variance, Table 5.9 shows that the standard deviations for H3, H2, and H4 are closer to each other ranging from 0.909, 0.926, and 0.935 respectively. That is the standard deviations are low which means that the scores for the themes are clustered close to the average score, i.e. students' scores on the various themes are somehow closer explaining approximate difficulty level and thus approximate performance alongside the themes.

Declaring facets for the item difficulty in the p(t:h) study design will serve as evidence of construct validity (structural or internal validity evidence). As mentioned in chapter three, statistical procedures can be implemented to assess the internal structure of an assessment, one aspect of Messick's validity evidence. Brown et al. (2019) have assumed that it is not obligatory that each item (theme) has the same value or duplicates the same difficulty level. They also stated that a mathematical model like this can determine students' ability and item diffulty generating a hiarachical scale of items from easy to difficult. In other words, G theory as a statistical model could difeferntiate between the themes at level of their difficulty, and this has been proved by means of calculating the mean scores and the standard deviations. This is another argument for the current test structural validity (Cardinet et al., 2010). Eventhough, the themes varied in terms difficulty they all underpin the overall test structure.

5.2.2.3. Descriptive Statistics for Tasks Nested within Themes in P(T:H) Design

In this section, descriptive statistics for tasks nested within themes is introduced to obtain information about the means, variances and standard deviations of the tasks nested within themes in order to verify item facility (and hence tasks difficulty). The results thus obtained are reported in Table 5.10.

H Themes	T :H Tasks within Themes	Mean	Variance	Std. Dev.
1	1	2.133	1.177	1.085
1	2	2.062	0.855	0.924
2	3	1.708	0.773	0.879
2	4	1.726	0.942	0.971
3	5	1.708	0.915	0.956
3	6	1.788	0.734	0.857
4	7	1.823	0.783	0.885
4	8	1.655	0.952	0.976

Table 5.10: Descriptive Statistics for Tasks Nested Within Themes in the p(t:h) Design

Table 5.10 shows that the performance mean scores of students in Task 1 (M= 2.133) and Task 2 (M= 2.062) nested within theme 1 are greater than the mean scores of their performance in other tasks nested with other themes. Comparing the mean scores of Task 1 and Task 2 with other tasks (from 3 to 8), they seem easier than the other tasks nested within themes (from 2 to 4). The estimated performance mean scores for Task 3 was (M= 1.708) and the mean score for Task 4 was (M=1.726) nested with theme 2, and the estimated mean score for Task 5 was (M= 1.708) and Task 6 (M= 1.788) nested within theme 3. In addition, the mean scores for Task 7 and Task 8 nested within theme 4 accounted for (M= 1.823) and (M=1.655) respectively.

As for the variances in the scores of the tasks nested with themes, these seem somewhat closer, ranging between (0.734 and 1.177) although Task 1 nested within Theme 1 had the largest variance (S^2= 1.177, and the least variance was in Task 6 nested with theme 3 which reached S^2=0.734). In consequence, the tasks can be classified in terms of difficulty, from easy to difficult, as follows: T 1, T 2, T 7, T 6, T 4, T 3, T 5, and T 8 (note that Tasks 5 and 3 are of equal difficulty).

In effect, the previous section examined item facility for themes, this section investigates item facility for tasks and tasks nested within themes interaction effects. The mathematical procedures in Table 5.10 could successfully differentiate the tasks according to their level of difficulty, and then could classify them accordingly. Thus, the measurement procedure argues to be valid (and have structural validity) as it could differentiate and classify the eight tasks in terms of difficulty (Cardinet et al., 2010).

5.2.2.4. Analysis of Variance

The following ANOVA table sums up the individual score variations for P, H, T:H, PH, PT:H, e relevant to the second P(T:H) design. As mentioned previously, the ANOVA table eventhough important may not be necessarily interpretable for two reasons: first because the major study focus is not put on individual scores but on total scores and this stable represents individual score variations. Second, if interpreted in terms of dataset, the next table would be a replication of ANOVA results. Only an overall description would be provided for the ANOVA results as it was put forward in the first design.

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	533.635	112	4.765	0.546	0.553	0.553	58.9	0.079
H	22.429	3	7.476	0.030	0.030	0.030	3.2	0.021
T:H	2.257	4	0.564	0.002	0.002	0.002	0.2	0.003
PH	134.321	336	0.400	0.046	0.046	0.046	4.9	0.018
PT:H, e	137.743	448	0.307	0.307	0.307	0.307	32.7	0.020
Total	830.385	903					100%	

Table 5.11: ANOVA Table for the Item Difficulty Study

Note. Item facility refers to tasks nested within themes facility.

Table 5.11 displays different values for different effects depicted in the three columns (random, mixed and corrected) of estimated variance components. Differences in values between the “Random” and “Mixed” columns are explained by Cardinet et al. (2010) by a mixed model design, in which the facet features are fixed. However, if the

facets of persons, tasks, and themes were all random, the random model values, rather than mixed values are to be used in the G analysis (view the first design ANOVA). After the application of the Whimbey's correction (See Chapter 3) to the variance components that are associated with effects of the fixed facet attributes, the results of the "Corrected" column and the "Mixed" become different.

In brief, differences in values between persons were substantial (58.9 % of the total variance). The perusing large residual component (32.7 % of the total variance) indicates large differences in the relative standing of persons on different tasks, large unmeasured variation, or both.

5.2.2.5 Measurement Design

G study information provided in Table 5.12 below summarizes the estimated variance components, error variances, and generalizability coefficients for the P/TH current measurement and specifically for the average of the four themes and for each theme individually.

Source of variance	Differentiation variance	Source of variance	Relative error variance	% Relatve	Absolute error variance	% Absolute
P	0.553		.....		.....	
	.....	H	.....		0.003	6.5
	.....	T:H	.....		0.000	0.6
	.....	PH	0.005	10.7	0.005	10.0
	.....	PT:H,e	0.038	89.3	0.038	82.9
Sum of Variances	0.553		0.043	100 %	0.046	100%
Standard Deviation	0.744		Relative SE: 0.207		Absolute SE: 0.215	

Coef_G Relative	0.93
Coef_G absolute	0.92

Table 5.12: G Study Table for the Fraction Themes Study (Measurement Design P/TH)

Note. Grand mean for levels used: 1.825, Variance error of the mean for levels used: 0.009, Standard error of the grand mean: 0.093.

As can be seen in Table 5.12 above, using one judge and four themes (eight tasks nested within four themes) to measure students' behaviour; students' productive depth of vocabulary knowledge yielded an acceptably high generalizability and dependability coefficients: 0.93 and 0.92 respectively. One justification for high G coefficients is the fact that the variability due to raters and themes are both not substantial, thus it is not necessary to have several other raters and themes to achieve an optimal reliability.

The ANOVA data set illustrates that the largest source of variance refers to the PT:H,E, that is confounded with any as yet unidentified sources of variance and with random error, by 89.3% for relative error variance and 82.9 for absolute error variance. The second substantial source of error is denoted by PH interaction by 10.7% for relative error variance and by 10.0 % for absolute error variance. As to H variance, it contributes to absolute error variance by 6.5%, and by 0.6% for T:H . These values are negligible compared to the potential error variances of PT:H, e and PH. This might suggest that the main effect for themes on students' performance is relatively low.

It is important to note that since themes is a fixed facet, the themes main effect cannot (and if so only slightly) contribute neither to relative or to absolute error variance, and the same can be said for the direct interaction relationships involving this facet. Subsequently, as Table 39 indicates, the purpose of this nesting relationship aims to check how well the fraction tasks/themes can be located relative to one another on a scale of difficulty or facility. The relevant interaction effects, therefore, are partitioned into three components: T:H, PH, PT:H,e. the interpretation for this measure depends on the absolute coefficient of reliability. The coefficient is 0.92 for dependability closer to

1, as denoted in Table 5.12, which means high reliability and accurate assessment precision for the present measurement situation.

5.2.3. Analysis and Presentation of the Third G Study Design Results

The aim of the third (G) study in the third partially-crossed design denoted **P×R(T:H)** is to obtain an estimate of components of universe score variance and error variance. It is also an attempt to check the ability of the present productive depth of vocabulary knowledge test to reliably differentiate between students and raters crossed with tasks nested within themes. The G study entails three divergent stages elicited in the observation and estimation designs where the measurement facets are described followed by descriptive statistics for the research main facets (raters, themes, and tasks nested within themes), analysis of variance, and the measurement design.

5.2.3.1. Setting up the P×R(T:H) G Study Design

The observation and estimation designs are changed compared to the two first designs. The present three facet p×r(t:h) partially crossed design involves four facets that feature in the G study. These are persons (p), raters (r), themes (H) and tasks within themes (T:H). The three facets of persons, raters and tasks are crossed because all the students responded to the same set of tasks and all the raters scored all students' performances, and hence they scored the same set of tasks. Since each two tasks of the eight tasks are embedded in one theme of the four themes, the facet tasks is nested within the facet themes, and the observation design is therefore PR(T:H) or (T:H)PR.

The PR(T:H) design is a three-facet partially crossed mixed design with H fixed. In the study of the students' productive depth of vocabulary knowledge, all the students responded to the eight tasks with varying themes: Ancient Civilization, Education in the World, Ethics in Business, and Feelings and Emotions. The themes differ from two tasks to the others categorized in dichotomous terms, so tasks are nested within themes. Since all the students answered all the tasks in all the themes, the design is partially crossed, with tasks (t) nested within themes (h) and both crossed with persons (p) and raters (R). We refer to this design as p×r(t:h). The themes observed in the G study were selected purposefully from the *"New Prospects'"* textbook content with no intention to generalize beyond them. In consequence, the theme facet is fittingly treated as fixed.

As far as the sampling status is concerned, the facets of students, raters and tasks are considered infinite random. This means that there exist an extremely large number of students and raters and even tasks that could be potentially selected to serve the research purposes. The theme facet is considered fixed (finite random), which means that there are only four themes that feature in the study purposefully chosen from the sixth themes of the targeted textbook. More information about G study major facets in play are described in Table 5.13.

Facet	Label	Levels	Univ.
Persons	P	113	INF
Raters	R	2	INF
Themes	H	4	6
Tasks withing themes	T:H	2	INF

Table 5.13: Observation and Estimation Designs for the Three Partially-Crossed PR(T:H) Design

Each facet comprising G analyses is identified by giving its full concept (persons, raters, themes, tasks nested in themes) in the first column entitled ‘Facet’. The second column indicates each facet by a label using the initial letter. The label ‘p’ for persons, the object of the measurement or the differentiation facet, ‘R’ for raters, ‘H’ for themes is not a letter initial because, for EduG it is not convenient to use ‘T’ for both tasks and themes in order not to have the same labels for different facets to avoid confusing results. The latter facets are determined as instrumentation facets. The third column headed “Levels” identifies the number of levels representing each facet in the data set. The facet P has 113 levels meaning the number of examinees; raters has two levels, which means that two raters scored the students’ performances and themes have four levels representing the topics of the four units in the intended textbook. The tasks being nested with themes identify the number of tasks (levels) embedded within themes; each two tasks among the eight tasks belong to the same theme of Ancient Civilization, Education in the World, Ethics in Business, and finally Feelings and Emotions. This sort of information, that is facet number level sampling and labelling comprise the full observation design.

As to the estimation design, within the same table (Table 5.13, p.259) a set of information about the size of every facet universe is provided in the column headed "universe"; the abbreviation INF or infinite describes the infinite random facets, namely persons, raters, and tasks that are sampled from infinite universes. As far as 'themes' is concerned, the theme is treated as a finite facet, as we did in the two previous G study designs, because the universe sizes of the fixed facets range between the number of observed levels and infinity (Cardinet et al., 2010). In practice, the targeted themes range between four among the sixth observed themes and infinite number of themes that have a potential to serve as a corpus in this research. Up to this point the sampling status of each facet is determined by levels and universe sizes.

5.2.3.2. Descriptive Statistics for Raters

The following Table 5.14 reports the descriptive statistics for the values of students' performance on vocabulary knowledge across raters.

R Raters	Mean	Variance	Std. Dev.
1	1.825	0.919	0.958
2	1.763	0.855	0.925

Table 5.14: Descriptive Statistics for Raters in the pr(t:h) Partially Crossed Design

Note. Grand mean: 1.794, Variance: 0.888, Standard Dev.: 0.942

Table 5.14 shows that, overall, the mean ratings for R1 is relatively high (1.82) when compared to R2 (1.75). Still, the mean weightings are approximate between R1 and R2, which means that they scored the eight tasks consistently; subsequently, consistency of test scores across raters is achieved. Also, as indicated in Table 41, the standard deviations are closer alongside the raters' judgments. Still, we notice some leniency for R1 when scoring performance.

In Design 1, the eight tasks/themes have arguably proved to be at acceptable level of difficulty and in Design 1 and Design 2, the raters were consistent in weighting the

tasks, nevertheless, this measurement situation entails relative measurement for both tasks and raters facets further needs to be accompanied by an exploration of the reliability of an absolute measurement in a G study.

5.2.3.3. Descriptive Statistics for Themes

After presenting the raters' judgements in the form of mean ratings of the students' performance on vocabulary, this section reports the descrtiptive statistics for the values obtained for the four themes in Table 5.15.

H Themes	Mean	Variance	Std. Dev.
1	2.013	0.942	0.971
2	1.679	0.842	0.917
3	1.788	0.840	0.916
4	1.697	0.857	0.926

Table 5.15: Descriptive Statistics for Themes in the Pr(t:h) Partially Crossed Design

Note. Grand mean: 1.794, Variance: 0.888, Standard Dev.: 0.942

Table 5.15 plainly shows that the theme mean scores and students' performance across themes, the main facet of design 3, are low when contrasted to the grand mean, except for H 1 (Ancient Civilizations) that surpassed the expected mean score accounting for 2.013 of the total mean score of 1.794, and H 3 by 1.788, which is closer to the grand mean of theme facet. This means that Design 3 arguably confirms Design 2 results indicating that H 1 (Task 1 and 2) followed by H3 (Task 5 and 6) to be easy compared to the other remaining two themes. This suggests that students performed well in these two themes (Ancient Civilizations and Ethics in Business) because they are easy.

Contrarily, the students' performance varied and seems weak both in H 2 and H4. These account for low mean scores for H 2 (by 1.679), H4 (1.697). Correspondingly, differences in theme mean scores especially H 4 and H1 sugest that H 1 is less complex than H 3, and H 3 is easier than H 4 and H 4 is rather easy than H 2; the later theme,

Education in the World, proved to be the most difficult theme for all the respondents. Hence, the measurement procedure could differentiate between themes in terms of their item difficulty. Consistently, themes can be ordered from easy to difficult starting with H1, then H 3, then H 4, and then H 2, confirming the study results yielded in Design 2.

As to the variance, the amount of variation existing among the students' performance across the four themes varies; ranging from very closer variation between H2 (0.842) and H3 (0.840), followed by approximate variation in H4 (0.857) and distant variation in H1 (0.942), overall, being compared to the total variance (0.888) and to the other themes.

Table 5.15 also shows that the standard deviations for H3, H2, and H4 are closer to each other ranging from 0.916, 0.917, and 0.926 respectively. That is, the standard deviations are low contrasted to the total standard deviation of 0.942, which means that the scores for the themes are clustered close to the average score, i.e. the students' scores on vocabulary performance on the various themes (H2, H3 and H4) are somehow closer indicating approximate difficulty level of the themes (and thus for the corresponding tasks) and thus approximate performance alongside the themes is reached. These results again replicate those of Design 2.

5.2.3.4. Descriptive Statistics for Tasks Nested Within Themes

The following quantitative display of results (Table 5.16) shows the descriptive statistics for tasks nested within themes, one of the major estimated variance components in Design 3.

H Themes	T Tasks withing themes	Mean	Variance	Std. Dev.
1	1	2.102	1.065	1.032
1	2	1.925	0.804	0.897
2	1	1.695	0.778	0.882
2	2	1.664	0.905	0.951
3	1	1.832	0.910	0.954
3	2	1.743	0.766	0.875
4	1	1.752	0.779	0.883
4	2	1.642	0.929	0.964

Table 5.16: Descriptive Statistics for Tasks Nested within Themes in the pr(t:h) Design

Table 5.16 reports the descriptive statistics for the tasks nested with themes. It clearly reveals that the students' performance mean scores in Task 1 (M= 2.102) and Task 2 (M= 1.925) nested within H1 are higher than the mean scores of their performance in other tasks nested with other themes. Across the four themes, when we compare the mean scores of Task 1 and Task 2 with other Tasks (from 3 to 8) are easier than other tasks nested within themes (from 2 to 4). The estimated performance mean scores for Task 3 was (M= 1.695) and the mean score for task 4 was (M=1.664) nested with theme 2, and the estimated mean score for task 5 was (M= 1.832) and task 6 (M= 1.743) nested within theme 3. Additionally, the mean scores for task 7 and task 8 nested within theme 4 accounted for (M= 1.752) and (M=1.642) respectively.

The variances in the scores of the tasks nested with themes were somewhat closer, ranging between (0.766 and 1.065) although task 1 nested within theme 1 had the largest variance (S^2= 1.065), and the least variance was in task 6 nested with theme 3 which reached (S^2=0.766). Correspondingly, the themes (and the tasks nested within) can be classified in terms of difficulty, from easy to difficult, as follows: H 1 (T 1+ T 2), T 7 (H4), T 6 (H3), H2 (T 4 , T 3), H3 (T 5), and H 4 (T 8). In consequence, the themes investigated in Design 3 varied in terms of difficulty level which means that the present

vocabulary test could reliably differentiate between students and raters crossed with tasks nested within themes.

5.2.3.5. Analysis of Variance

The ANOVA table (Table 5.17) contains information related to the students' individual scores and the total scores. So, as mentioned earlier in Design 1 and 2, there is no need to interpret the values because it will be a replication once the G study Table 5.18 will be interpreted in details. But still the ANOVA values are useful and can be indicative and helpful to better understand the G study table results.

Source	SS	Df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	894.585	112	7.987	0.440	0.445	0.445	49.2	0.066
R	1.735	1	1.735	0.000	0.000	0.000	0.0	0.002
H	31.969	3	10.656	0.020	0.020	0.020	2.2	0.015
T:H	5.916	4	1.479	-0.001	-0.001	-0.001	0.0	0.005
PR	50.890	112	0.454	0.034	0.034	0.034	3.8	0.008
PH	224.906	336	0.669	0.029	0.029	0.029	3.2	0.016
PT:H	245.084	448	0.547	0.184	0.184	0.184	20.4	0.019
RH	3.615	3	1.205	0.000	0.000	0.000	0.0	0.005
RT:H	5.111	4	1.278	0.010	0.010	0.010	1.1	0.007
PRH	61.760	336	0.184	0.003	0.003	0.003	0.3	0.009
PRT:H, e	79.889	448	0.178	0.178	0.178	0.178	19.7	0.012
Total	1605.460	1807					100%	

Table 5.17: Analysis of Variance for the Three Facet Partially Crossed pt(t:h) Design

Table 5.17 displays the G study measurement design P/RTH estimated variance components, error variances, and G coefficients. The sources of variability under study in this measurement situation involve: P, R, H, T:H, PR, PH, PT:H, RH, RT:H, PRH, PRT: H, confounded with unmeasured sources of error 'e'. In the EduG software

program, differentiation facets are declared to the left and instrumentation facets to the right of a slash (P/RTH), here persons are taken as a single differentiation facet and raters, tasks and themes are instrumentation facets. The objective of this application is to differentiate among test takers vocabulary performance.

5.2.3.6. Measurement Design

The measurement design or G study table (Table 5.18) below summarizes information related to the estimated variance components, error variances, and generalizability coefficients for the P/RTH current measurement and specifically for the average of the four themes and for each theme individually.

Source of variance	Differentiation variance	Source of variance	Relative error Variance	% Relative	Absolute error variance	% Absolute
P	0.445		.....		.....	
	.....	R	.....		0.000	0.2
	.....	H	.....		0.002	3.5
	.....	T:H	.....		(0.000)	0.0
	.....	PR	0.017	31.5	0.017	30.0
	.....	PH	0.003	5.4	0.003	5.1
	.....	PT:H	0.023	42.4	0.023	40.3
	.....	RH	.....		(0.000)	0.0
	.....	RT:H	.....		0.001	1.1
	.....	PRH	0.000	0.3	0.000	0.2
	.....	PRT:H,e	0.011	20.5	0.011	19.5
Sum of Variances	0.445		0.054	100%	0.057	100%
Standard Deviation	0.667		Relative SE: 0.233		Absolute SE: 0.239	
Coef_G relative			0.89			

Coef_G absolute	0.89

Table 5.18: G Study for P/RTH Design with Three Infinite Random Facets and One Fixed Facet

Note. Grand mean for levels used: 1.794, Variance error of the mean for levels used: 0.007, Standard error of the grand mean: 0.085.

The above table confirms that the two largest measurement error contributors, namely the interaction between students and tasks within themes (PT: H by 42.4%) and the interaction between students and raters (PR by 31.5%), then followed by the interaction between students raters and themes (Students by Raters by Tasks within themes PRT:H, e by 19.5% , all together account for over 93% of the total relative error variance, at five. Additionally, for the absolute measurement, there are eight contributors to the total error variance. PR 30%, PT: H 40.3%, and PRT: H, e 20.5% accounting for more than 89 % of the total absolute error variance. Amongst these, the most significant is the Student by Tasks interaction variation within Themes (PT: H) with a contribution of about 40 % less than half of the total error variance (89 %). When opposed to each other, there is a relatively small number between the relative and absolute total error variances (difference is 4%).

As to the G coefficients, Table 5.18 shows that only five potential variance components could contribute to the total error variance for relative measurement, on the other hand, eight potential sources are contributors to the total error variance for absolute measurement. The resulting Coef_G relative and Coef_G absolute is 0.89 which are reasonably high, at 0.89. As a matter of fact, the relative G coefficient for the current productive depth of vocabulary knowledge test has yielded a reliable estimation of the difference in themes levels set for the whole group of students being assessed by two different raters. Besides, this confirms that the test produced an index measurement reliability for absolute theme levels on each individual task. Similarity in the two G coefficients, in a great deal, is justified by the amount of error contributors that proved to be slightly different for both relative and absolute measurements. It is also due to the null T:H, RH; no inter-task variability within the four themes, which does not negatively

affect the coefficient of relative measurement, if inter-task variability within the four themes were high, the coefficient of absolute measurement would be low.

5.3. Optimization Design: D Studies

This section is an attempt to answer the D studies related questions associated with the different study designs. One of the possible improvements or optimization is to answer the upcoming questions:

1- What is the effect of decreasing the number of tasks designed to assess vocabulary performance on the generalizability and dependability of test scores?

2- What is the effect of decreasing the number of raters on the generalizability and dependability of test scores?

5.3.1. Optimizing Measurement Precision

The G study analyses have revealed so far the different sources of variability and indicated the relative impact, whether low or high, they have on error variance. Based on this key information, we have made some practical decisions on how to improve the current measurement. The EduG system allowed us to determine an optimal observation design for the present measurement situation.

The objective of D studies is to use sources of variability obtained in G studies in order to design a measurement procedure that can minimize measurement error (Bachman, 2004; Gerbil, 2009). In practice, this section is an attempt to answer question 1 and question 2 above by providing a set of scenarios deploying various combinations of tasks, and raters, as they represent major sources of variance in the p×t×r design affacting the study scores. In essence, in the D studies different combinations containing different sets of tasks and raters are suggested to determine how well changing the number of tasks and raters affects the measurement precision.

Study approximations are made according to this principle: “where possible, increase the numbers of observed levels of the greater contributors to error variance and, if it contributes to cost-effectiveness, decrease the numbers of observed levels of those instrumentation facets that contribute little to the error variance” (Cardinet et al., 2010). Accordingly, we made some successive approximations until we arrived at

accurate optimum design producing acceptable degree of reliability. These approximations have been made in accordance with the numbers of observed levels just for instrumentation facets (tasks and raters), but not for the differentiation facet (students) in order to check the potential impact due to change on the index of measurement reliability. This optimization facility made it possible to reduce the number of observed levels for those facets of tasks and raters that contribute to the error variance. Note that the task and rater main effects contributed little but when interacted with other facets they largely contributed to error variance and thus affected the measurement precision and reliability.

To optimize the measurement precision, four D studies were conducted in two phases:

Step One: as to the measurement design *P/TR* discussed earlier, we altered the number of observed levels beginning with the number of tasks in order to answer the following sub-question: *How may tasks would be needed to maximize an index of reliability?*

Step Two: and then we modified the number of raters to obtain an optimal reliability to answer the question: *How many raters would be required to maximize the reliability in an in-depth productive vocabulary test?*

Most importantly, the size of the sample (113) and the universe is infinite for P, the dependent variable, both cases remain the same. The EduG work screen allows the user to have five options for facet level numbers (see Appendix G). Table below (Table 5.19, p. 269) depicts the optimization results whereby the original acompnied by combination values, the resulting G and phi coefficients, variance and error estimates are displayed.

5.3.2. Decision Studies

This section is an attempt to answer the D studies questions which are related to an investigation of the effect of decreasing the number of tasks and raters on the generalizability and dependability of test scores obtained from the vocabulary performance test. In order to examine how much impact the number of tasks and raters might have on maximizing or minimizing score reliability for the vocabulary test scores,

in this section the independent tasks variable will be investigated then followed by rater independent variable. Implicit in G theory is a univariate analysis that focuses on each variable independently. In the upcoming table (Table 5.19) the generalizability coefficient and dependability coefficient together with their rounded values, error variances for both relative and absolute measurement are calculated to gain a clear image of the score reliability across tasks within different combinations. To this end, a number of D studies have been conducted within the following scenarios:

5.3.2.1. Optimization 1: Decreasing the Number of Tasks

D study 1: six tasks and two raters

D study 2: four tasks and two raters

D study 3: three tasks and two raters

D study 4: two tasks and two raters

The results thus obtained are detailed in Table 5.19 to answer question 1.

	G-study		Option 1		Option 2		Option 3		Option 4	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	113	INF	113	INF	113	INF	113	INF	113	INF
T	8	INF	6	INF	4	INF	3	INF	2	INF
R	2	INF	2	INF	2	INF	2	INF	2	INF
Observ.		1808		1356		904		678		452
Coef_G rel.		0.895		0.873		0.833		0.797		0.732
Rounded		0.89		0.87		0.83		0.80		0.73
Coef_G abs.		0.890		0.867		0.825		0.787		0.720
Rounded		0.89		0.87		0.83		0.79		0.72
Rel. Err. Var.		0.053		0.066		0.090		0.115		0.165
Rel. Std. Err. of M.		0.230		0.256		0.301		0.339		0.406
Abs. Err. Var.		0.056		0.069		0.096		0.122		0.175
Abs. Std. Err. of M.		0.236		0.263		0.309		0.349		0.418

Table 5.19: Optimization 1: Reduction of Tasks with Constant Raters

The above table shows the first optimizing measurement precision for the first design whereby the task facet is reduced. The relative error variance is often used in norm-referenced interpretations where test stakeholders intend to rank order test takers. The results indicated that the relative error variance of the tasks have a **slightly** higher value whenever the tasks are reduced. For example, the relative error variance of having eight tasks and one rater is 0.053; when having six tasks and one rater the relative error variance is 0.066; and 0.099 when the tasks are reduced to three it reached 0.115, and finally it is 0.165 with the minumun of two tasks. It is quite remarkable, as it is demonstrated in the table above that decreasing the number of tasks from eight to two tasks considerably resulted in increasing the relative error variance in the present measurement situation. Thus, the largest reduction of error variance occurred in the first scenario of eight tasks with two raters, which is the original plan for the G study. The results as such suggest that decreasing the number of tasks rather than the number of raters yielded a substantially increased error of measurement. Changing or reducing the number of tasks and keeping the same rater rate negatively, but slightly, affected the actual measurement precision.

The absolute error variance, on the other hand, is used particularly in criterion-related interpretation where the main emphasis is put on the examinees' performance with relation to a specific standard but not to elicit the relative standing of examinee's against a continuum (i.e. the relative standing of examinees against each other). The D study results displayed that the absolute error variance is slightly higher when compared to the relative error variance obtained throughout the different scenarios. The absolute error variance in the different combinations is becoming rather high. That is, when the tasks are reduced to six the absolute error was 0.056; when reduced to four it was 0.069; and when reduced to three and then to two it becomes 0.122 and 0.175 respectively. Reduction of tasks variable rather than raters variable, therefore, increases the absolute error variance, in essence, as the relative error variance.

The generalizability coefficient in Brennan's (2001) terms is analogous to the CTT

reliability coefficient. The results in Table 5.19 demonstrate that the G coefficient had relatively high to acceptable degrees across all the D studies except for the last two scenarios of reducing the amount of tasks to two keeping the number of raters constant (n=2). For instance, as shown in the table, the G coefficient was 0.873 and 0.833 when the tasks are reduced to six and four respectively and was dropped to 0.797 and 0.732 when the tasks were decreased to three and two respectively. In the initial G study it was 0.895 which is highly acceptable to draw study generalizations. Hence, the most acceptable improvement was achieved when the tasks were reduced to six and the minimum G coefficient resulted from decreasing the number of tasks to four. But when the tasks were decreased to three and two, the G coefficient was lower than 0.80 (by 0.73 and 0.70 respectively) the value that cannot be used to generalize the results in performance-based assessment. Hence, we need to reduce the number of tasks up to four with two raters in order to obtain an acceptable degree of generalizability of scores.

As far as the phi coefficient, equated to dependability coefficient, is concerned, it ranges between high and low values across the successive approximations suggested in the D studies. The accurate optimum design producing acceptable degree of reliability lies in reducing the tasks to four keeping, of course, raters condition constant (R=2). However when reduced to three and two the values produced for phi coefficient are below 0.80 (by 0.787, and 0.720) the convention scale for which researchers can make their generalizations. In consequence, in order to produce an acceptable dependability index of scores, the number of tasks should be decreased up to four in this particular case.

5.3.2.2. Optimization 2: Decreasing the Number of Tasks

Since the EduG software does not provide the user with more than five approximations (columns), we were in a position to conduct another optimization design where the tasks are also reduced to seven, five, three ad finally two in addion to reduction of rater from two to one is conducted. The resulting four D studies are illustrated as under and the final results of reduction are summarized in Table 5.20 (p.272).

Decreasing the number of tasks

D study 1: seven tasks and two raters

D study 2: five tasks and two raters

D study 3: three tasks and two raters

D study 4: one task and two raters

The D studies involving three tasks and two raters and two tasks and two raters, written in bold, will be eliminated from the analysis and interpretation in this section because they have already been tackled in optimization 1. Options 3 and 4 are kept in the table of optimization 2 to foster understanding of how G values are different through the five options.

	G-study		Option 1		Option 2		**Option 3**		**Option 4**		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	**Lev.**	**Univ.**	**Lev.**	**Univ.**	Lev.	Univ.
P	113	INF	113	INF	113	INF	**113**	**INF**	**113**	**INF**	113	INF
T	8	INF	7	INF	5	INF	**3**	**INF**	**2**	**INF**	1	INF
R	2	INF	2	INF	2	INF	**2**	**INF**	**2**	**INF**	2	INF
Observ.	1808		1582		1130		**678**		**452**		226	
Coef_G rel.	0.895		0.885		0.857		**0.797**		**0.732**		0.590	
rounded	0.89		0.89		0.86		**0.80**		**0.73**		0.59	
Coef_G abs.	0.890		0.880		0.850		**0.787**		**0.720**		0.575	
rounded	0.89		0.88		0.85		**0.79**		**0.72**		0.57	
Rel. Err. Var.	0.053		0.058		0.075		**0.115**		**0.165**		0.314	
Rel. Std. Err. of M.	0.230		0.242		0.275		**0.339**		**0.406**		0.560	
Abs. Err. Var.	0.056		0.062		0.080		**0.122**		**0.175**		0.334	
Abs. Std. Err. of M.	0.236		0.248		0.282		**0.349**		**0.418**		0.578	

Table 5.20: Optimization 2: Reduction of Tasks with Constant Raters

We discarded options 3 and 4 with tasks reduced to 2 and 3 in this table since they

have been discussed in Table 5.19 above. The two options were inserted because the EduG application conditions five options to compute the D studies variance components, error variances and G coefficients. Note that the relative and absolute G coefficients are similar to those obtained in the *PTR* G studies as reported in Table 5.19. However, when the number of tasks was reduced to seven the Coef_G rel reached 0.885 which is approximately equal to eight tasks in the initial design before optimization design occurred. The results in Table 47 suggest that it is possible to have an optimal reliability by designing just seven tasks instead of eight. As for option 3 where the tasks were decreased to five, the G coefficient obtained was 0.857, but was not acceptable when the tasks altered to one. Thus having one task in the test would not yield an acceptable reliability index (0.590). The coefficient of absolute measurement was slightly low across the different options as opposed to the coefficient of relative measurement (Option 1 and Option 2 resulted in 0.880 and 0.850 phi coefficients respectively). Like the G coefficient, the phi coefficient was low (0.575) in option 5.

Both coefficients are slightly different and remain somewhat stable in the first three options, which means that the generalizability and dependability of scores across the options were acceptable.

The D studies noticeably indicated that the relative and absolute error variances increased throughout the five options. The relative error variance ranges from 0.053 in option 1 incorporating eight tasks, 0.085 in option 2 with seven tasks, 0.075 in option 3 integrating five tasks and 0.314 in the final option with just one task. It is clear that the relative error variances increased whenever the number of tasks was reduced.

Changing the number of tasks has also affected the measurement precision. In the D study analysis upon removal of tasks the absolute error variances also increased by decreasing the number of tasks and keeping constant raters variable. As shown in Table 5.20 the absolute error variances varied between 0.056, 0.062, 0.080, and 0.334 across the options suggested for reduction of some tasks from eight to seven to five to one respectively.

5.3.2.3. Optimization 3: Decreasing the Number of Both Tasks and Raters

In similar veins, the optimization design for the PTR design goes through the

following five D studies senarios. The results obtained are stated in Table 5.21 coming next.

D study 1: eight tasks and one rater

D study 2: six tasks and one rater

D study 3: four tasks and one rater

D study 4: three tasks and one rater

D study 5: one task and one rater

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	113	INF	113	INF	113	INF	113	INF	113	INF	113	INF
T	8	INF	8	INF	6	INF	4	INF	3	INF	1	INF
R	2	INF	1	INF	1	INF	1	INF	1	INF	1	INF
Observ.	1808		904		678		452		339		113	
Coef_Grel.	0.895		0.850		0.825		0.779		0.738		0.520	
Rounded	0.89		0.85		0.82		0.78		0.74		0.52	
Coef_Gabs.	0.890		0.844		0.818		0.770		0.728		0.505	
Rounded	0.89		0.84		0.82		0.77		0.73		0.51	
Rel. Err. Var.	0.053		0.080		0.096		0.128		0.160		0.417	
Rel. Std. Err. of M.	0.230		0.282		0.310		0.358		0.400		0.646	
Abs. Err. Var.	0.056		0.083		0.101		0.135		0.169		0.442	
Abs. Std. Err. of M.	0.236		0.289		0.317		0.367		0.411		0.665	

Table 5.21*:* Optimization 3: Reduction of Tasks and Raters

A review of Table 5.21 indicates that, when the number of raters is reduced to one and tasks reduced to six, four, three, and one successively, the Generalizability and dependability coefficients maintain somewhat stable across combinations. For example, when the tasks held constant (T=8) in the first option and the number of raters reduced

from two to one, the G coefficient for relative measurement was 0.850 and 0.844 for the absolute coefficient of measurement with only 0.45 and 0.51 respective differences when compared to the G coefficients initially produced in the first G study. However, error variances attributed to tasks and raters produced acceptable index of reliability. For relative measurement the G coefficients range between 0.779, 0.738 and 0.520; and for absolute interpretations the absolute coefficient spread from 0.770, 0.728, and 0.505. Therefore, the produced G and phi coefficients were relatively stretch from high to acceptable with slight differences between the coefficient of relative measurement and coefficient of absolute measurement.

The relative error variances are relatively small when reduction of raters was held to one. The largest error variance reached 0.417 with one task and one rater in D study 5; with one task and one rater and the smallest error variance was 0.080 for D study 1 with eight tasks and one rater. The absolute error variances were remarkably low across the suggested approximation. For instance, in option 1 the absolute error of measurement was 0.083, in option two was 0.101 and in option 3, 0.135, and in the remaining two options it was 0.169 and 0.442. This explains the fact that the number of tasks and raters affects the estimate of the generalizability coefficient up to about one task and one rater. Consequently, the use of one rater instead of two would result in a very little effect on the reliability of scores as Table 5.21demonstrates with coefficients relatively high or below 0.80 considered acceptable or extremely acceptable.

5.3.2.4. Optimization 4: Decreasing the Number of Tasks and Raters

Four approximations have been suggested to investigate the effects of decreasing the number of tasks and raters along the four D studies.

D study 1: seven tasks and one rater

D study 2: five tasks and one rater

D study 3: three tasks and one rater

D study 4: two tasks and one rater

The results thus obtained are reported in the Table 5.22 underneath.

	G-study		Option 1		Option 2		Option 3		Option 4	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	113	INF	113	INF	113	INF	113	INF	113	INF
T	8	INF	7	INF	5	INF	3	INF	2	INF
R	2	INF	1	INF	1	INF	1	INF	1	INF
Observ.		1808		791		565		339		226
Coef_G rel.		0.895		0.839		0.806		0.738		0.668
Rounded		0.89		0.84		0.81		0.74		0.67
Coef_G abs.		0.890		0.832		0.798		0.728		0.655
Rounded		0.89		0.83		0.80		0.73		0.66
Rel. Err. Var.		0.053		0.087		0.109		0.160		0.224
Rel. Std. Err. of M.		0.230		0.294		0.330		0.400		0.474
Abs. Err. Var.		0.056		0.091		0.114		0.169		0.237
Abs. Std. Err. of M.		0.236		0.301		0.338		0.411		0.487

Table 5.22: Optimization 4: Decreasing the Number of Tasks and Raters

It is apparent from examining Table 5.22 that the Combinations 1 and 2 display the results for tasks and raters reduced with an initial reliability of 0.839 and 0.806 for relative measurement and 0.832 and 0.798 for dependability interpretations, followed by 0.738 in option 3 and 0.668 in option 4, accompanied by respective 0.728 and 0.655 for phi coefficients. As shown in the table, in all the combinations the generalizability estimates decreased as the number of raters decreased to one. As a consequence, reduction of tasks from eight to seven to five to three to two lowered the generalizability coefficients from 0.895 to 0.668 in general. As to the phi coefficients, they range between 0.832, 0.798, 0.728 and 0.655 that are noticeably approaching the G

coefficients of .80.

The optimization studies and procedures reported in Table 5.22 for each of the aforementioned four D studies account for the largest error variance occurred in option 4 by 0.224 for relative measurement and 0.237 for absolute measurement and the lowest error variance was in option one by 0.087 and 0.091 for both relative and absolute error variances respectively. This suggests that the error variance increases whenever the number of tasks is changed from eight to two and when the number of raters is reduced to one, and this in turn, affects the generalizability estimates or the reliability index; the more error variance increased the lower G coefficients would be obtained and vice versa. The resulting high and low temporal reliability explains no sharp increase in error variances whilst modifications at the level of measurement conditions occurred (at facet levels sampling).

What is worth noting, here, is that the findings for these reductions were slightly dissimilar to those obtained in all tables before as the reliability index was somewhat stable across D studies combinations and scenarios.

Conclusion

This chapter has been devoted to the analysis and presentation of quantitative data obtained from the three G study fully crossed, partially-nested and partially-crossed designs, and from the four D studies or optimization designs. Data was elicited through numerical statistical tables derived from the EduG software workscreen prcedures. The upcoming chapter tackles the discussion and interpretation of the findings. The quantitative data obtained from Design 1, Design 2, and Design 3 are compared. These data accompanied with other results of optimization design combinations and scenarios will be used to answer the research questions constructed at the commencement of this research.

CHAPTER SIX

DISCUSSION, INTERPRETATION OF THE FINDINGS AND IMPLICATIONS

Introduction

In this chapter, we seek to discuss and interpret the research findings, obtained so far from the generalizability analyses, displayed in the previous chapter regarding both the G studies and D studies. We endeavor to answer the upcoming research questions relevant to investigating the relative impact of tasks, raters and themes on the generalizability and dependability of the observed test scores obtained from the vocabulary performance assessment:

Part I: G studies related questions:

***Design 1:* Two-facet fully crossed p×t×r design**

RQ1: What is the relative effect of tasks and raters on the generalizability (for relative decision) and dependability (for absolute decision) of the test scores obtained from the vocabulary performance assessment?

***Design 2:* Two-facet partially nested P (t:h) or p(h:t) design:**

RQ2: What is the relative effect of tasks and themes on the generalizability and dependability of the scores obtained from the vocabulary performance test?

***Design 3*: Three-facet partially nested Pr(t:h) design:**

RQ3: What is the relative effect of raters, tasks and themes on the generalizability and dependability of the scores obtained from the vocabulary performance test?

Part II: D studies related questions:

RQ4: What is the effect of decreasing the number of tasks designed to assess the vocabulary performance on the generalizability and dependability of the test scores?

RQ5: What is the effect of decreasing the number of raters on the generalizability and dependability of the test scores?

Part III: Collecting validity evidence

RQ6: What is the relative effect of tasks, raters and themes on the construct validity of the vocabulary performance assessment?

This chapter is also an attempt to draw some implications based on the research findings and discussions. These implications are consolidated by previous literature, especially addressed to assessment specialists, be they educators, researchers and stakeholders. It recommends, among other things, to apply the principles of G theory to the assessment of tests in terms of their psychometric properties, including reliability and construct validity.

This study is descriptive in design wherein G theory method was implemented to carry out the generalizability analyses via ANOVA procedures. The quantitative data was gathered by the in-depth productive vocabulary knowledge test assigned to first year university EFL students at ENSB. The EduG software program was utilized to analyze the data which were presented via generalizability statistical tables. This software package fits the present research work, because it enabled us to determine the various sources of variance and instability that has strong effects on observations (i.e. on the generalizability and dependability of the obsereved test scores). The software was also able to determine the impact of altering the observation design intended to eliminate the principal contributions (largest sources of variability) to the measurement error variances. In consequence, it enabled us to estimate the reliability of the measuring instrument under study, taking into account the estimation of variance components having an effect on the score generalizability. The overall aim of using the EduG procedure is to assess the assessment and increase the measurement situation's optimum reliability index.

This statistical program had arguably proved to be effective in terms of calculating, triangulating and presenting the results gained from the generalizability analyses. The researcher, thus, will interpret and discuss these results along the following sections. The associated research questions will also be interpretively and sequentially answered in different sections.

6.1. Relative Effects of Tasks and Raters on the Productive Depth of Vocabulary Knowledge Scores

This section seeks to answer RQ1 associated with Design 1, the two-facet fully crossed p×t×r design, which is formulated as follows:

RQ1: What is the relative effect of tasks and raters on the generalizability (for relative decision) and dependability (for absolute decision) of the test scores obtained from the vocabulary performance assessment?

In the two-facet person-by-task-by-rater fully-crossed design (p×t×r), the results produced so far by G analyses revealed that, of the seven interaction effects, **the largest source** of variability that had an increasingly deep impact on the generalizability and dependability of the scores obtained from the in-depth productive vocabulary knowledge test refers to the variance component for **person-by-task interaction (PT)** for both, the relative measurement (by 49.7%) and absolute measurement (by 47.2%). This indicates task-to-task variation in the performance of individual students. PT interaction effects reflect within-task instability in the examinees' individual performance as every student completed the task in a single time, which means that tasks have not been replicated or re-tested (Cronbach et al., 1997).

The results obtained agree with those of Lee and Kantor (2007). Using a multitask writing measure, they confirm that of the total variance sources of variance highly affecting performance are associated with person-by-task variance and task variance, demonstrating that the relative ranking of examinees varies considerably across different tasks. Task sampling variability (person × task) has been approved to be a major source of measurement error in various studies, either in the writing area using task writing performance as in Gao and Brennan (2001), Huang (2009, 2012), and Gebril (2010), or in mathematics performance assessment illustrated in Lane's et al.(1996) G study that reported in a p × t × r crossed design that person-task interaction accounts for the largest proportion of the total variability (by 54%, 55%, 41%, and 62%) for rater pairs 1, 2, 3, and 4 respectively. These results, together with the current findings, are consistent with the findings of other studies in mathematics performance

assessments (Shavelson, et al., 1993; Hébert, 2006; McBee & Barens, 1998) and in science performance (Shavelson et al., 1993; Webb et al., 2000).

One more finding shows that the main effect for tasks accounted for 3.3 % resulting in inter-task consistency, and the main effect for raters was 0.5 % leading to inter-rater reliability of the total variance. These main effects were minimal. These results, in conjunction with Lee's and Kantor (2007) study indicated that, there existed some difference in the task difficulty among the tasks administered in the writing performance. The person-by-task variance explains that the greatest source of variability in the students' test performance was attributed to the differences among students' vocabulary knowledge and ability measured via the writing tasks. It also indicates differences in the students' mean performances across the tasks, that is, the mean performance of every individual student varies from one task to the other and this, of course, is a matter of the relative simplicity and difficulty of the tasks devoted for completion.

As for the **low main effect for raters,** Design 1 demonstrates that "there were no marked differences in stringency or leniency between the assessors that might have distorted the assessment system" (José-Luis Menéndez-Varela & Eva Gregori-Giralt, 2017, p. 6). The between-rater variance (0.5%) is relatively very low, which means that the raters were consistent in their weighing of the students' performance on the quality of word knowledge. As such, inter-rater consistency has been arguably proved in the application of the scoring rubrics and providing appropriate subjective judgments on the individual performance or students' responses. These results are consistent with those of Polat and Turhan (2021), who concluded that "the stringency and leniency levels of the raters' scores to students' performances do not differ significantly"(p.3351). The researchers investigated students' speaking skill via open-ended tasks and they could find the main rater effect to contribute to the total variance only by 0.3%.

The second source of error variance depicted in the **students-by-rater interaction** (PR) accounted for 29.8% for relative measurement and 28.3% for absolute measurement. This explains how the raters slightly differed in the interpretation and

application of the rubric components for each test taker in PTR pattern; being influenced by the students' performance, they deviate from the usual interpretation of the scoring rubrics and, hence, provide distinct judgments per each person, as a result of the way they perceive the students' performance and ability. This finding is supported in other research. In his study on the accuracy and validity of the writing scores assigned to secondary school ESL students in the provincial English examinations across three years, Huang (2012), indicted that Person-by-rating (PR) was the second largest variance component for 2001, 2002, and 2003 (by 23.1%, 22.4%, and 29.0% of the total variance, respectively). This relatively high proportion of PR contribution to the total of measurement error suggests that there was inconsistency in rating severity or leniency for ESL students. This finding, however, disagrees with what is being reported in other research. For example, the person-by-rater (PR) effect accounts for only 3.5% in the independent tasks and 3.0 % in the integrated tasks exploring writing scores (Gebril, 2010). This measurement error explains that the raters are roughly dissimilar in how they scored the paragraph responses across the respondents. Put simply, the ratings change over the students; from one student to the other. The task-by-Rater (TR) interaction effects on scoring the students' short constructed paragraphs were not significant (accounting only 1.1% of the total variance). This evidence is supportive of internal consistency.

The third variance component having an increased effect on the actual measurement precision is the **students-by-task-by-rater interaction confounded with all unidentified sources of variance (ptr, e)**. In the actual measurement situation, **ptr, e** proportion of the total variance was 20.5% for relative measurement and 19.5 % for absolute measurement. These findings replicate the results obtained by Lee and Cantor (2007) who came to the conclusion that, the second largest variance component was that associated with the person-by-task-by-rater interaction plus **e**. It differentiated error [σ2 (ptr, undifferentiated)] in the p × t × r design accounting for 20.5% of the total variance. Lee and Cantor propose that "examinees were rank-ordered less consistently across different task-by-rater pairs" (p. 373). In Gebril (2005), the value for ptr component obtained for both the independent and integrated task types accounted for (40% and 41%, respectively). The difference, however, lies in the order of this source

of measurement error. In Gebril's case presents the first largest component, for Lee and Cantor, it proved to be the second, and in the current measurement, it stands for the third variance. Depending on the nature and number of tasks and the type of designs applied, these studies yielded different percentages for the relative effects that the person-by-rater-by-task interaction might have on the measurement error contribution.

As to the generalizability (G) and dependability (phi) coefficients for the present two types of measurement, respectively for both the relative and absolute, the G coefficients values in the G study within the first design was **0.89**. The G coefficients values for both the relative and absolute measurements go hand in hand with the previous research on performance-based writing assessment. Huang (2012), for example, could find the G coefficients of 0.85, 0.89, and 0.88 for the writing scores in the years 2001, 2002, and 2003 respectively.

The G coefficients values for both the relative and absolute measurements are approximate to the previous research on performance-based assessment. To illustrate, in assessing the speaking skill, Polat and Turhan (2021) obtained a high G coefficient of 0.85 but low phi coefficient of 0.78 in the nested design. The relative G-coefficient value was .86 and for the absolute G-coefficient was .85 (Khodi, 2021). Webb et al. (2000) could also reach a 0.85 dependability coefficient.

In the meanwhile, these values disagree with other studies that did not reach the standard criterion set for generalizability in performance assessments. They concluded less acceptable G coefficients with the value less than 0.80. (e.g.; Shavelson et al., 1993; Ruiz-Primo et al., 1993; McBee & Barens, 1998; Gao & Brennan, 2001; Polat & Turhan, 2021 (G: 0.79, Phi: 0.72)).

As such, the value in this current case is extremely acceptable with the current measurement conditions (number of students, tasks, raters and themes), as values of at least 0.80 are capable of making generalizations of test scores in performance-based assessment. According to Cardinet et al., (2010), G-coefficients beyond 0.80 explain a satisfactory reliability for certain measurement situations. It informs us of the degree to which stability of the students' scores across tasks have been achieved. As a matter of fact, the universe score distribution persists stable across the tasks. In chapter three, we

confirmed that the values set for the G coefficients range between 0 and 1. That is, the test was more reliable because the value was higher. This concludes that the test is a reliable procedure used to get the average of vocabulary performance for each of the eight tasks' score precision.

Noticeably, the previous studies indicated differences in the findings of G coefficients due to many reasons: differences in the variance components contributions to measurement error, differences in the nature and number of tasks, differences in the number of facets levels (e.g., two raters or more), … etc. The G coefficients also varied in the G theory research endeavor depending on the design implemented; whether random effects in crossed designs or mixed effects in nested designs.

6.2. Relative Effects of Tasks and Themes on the Productive Depth of Vocabulary Knowledge Scores

This section aims to answer RQ2 associated with Design 2, the two-facet partially nested p(t:h) or p(h:t) design**,** which is constructed as follows:

RQ2: What is the relative effect of tasks and themes on the generalizability and dependability of the scores obtained from the vocabulary performance test?

In the second two-facet partially nested p×(h:t) study design, the ANOVA data set showed that the largest source of variance was attributable to **the residual component pt:h, e** by 89.3% for the relative error variance and 82.9 for the absolute error variance other than the person-task interaction obtained in the first random crossed effects design. Thee G analyses revealed that the two-way interaction nested within themes facet, plus undifferentiated error variances (PT:H,e), was substantially overestimated in Design 2, when compared with Design 1 and Design 3 analyses. This explains that more random errors are included in the measurement procedure while using the design in which the variance components of persons and tasks are nested in the themes compared to the full factorial crossed model. Keeping the same facets of measurement of Design1, the facet of themes was added and the design was nested (from P×T×R design to P×(H:T) design). In consequence, the source of the residual effect variability has the highest variance percentage in the second partially nested design.

Previous studies have also affirmed that the residual components arguably seemed to be the largest variance component, particularly, in language testing studies applying G theory, dealing with performance assessments of the ESL speaking skill (e.g., Lynch & McNamara, 1998; Polat & Turhan, 2021), and EFL writing performance (Gebril, 2010). In a similar vein, in an attempt to investigate the variability and reliability of holistic scores of eighty argumentative EFL essays written on two distinct themes by tertiary level Turkish students (note that themes here are not treated as facets of measurement), Sari and Han (2022) assert that the residual component proved to be the greatest source of variance by 40.1 % of the total variance, affecting scores due to interaction between human raters and students responses together with systematic and unsystematic error sources. In Huang (2012), however, the residual yielded the second largest variance component.

The residual component (**pt:h, e)** involves the variability due to the interaction between students, tasks within themes, and other unexplained systematic and unsystematic sources of error. The large variability in the residual effect was a significant result because, thus far, no studies have yielded a similar degree of variability, especially that themes as facet of measurement error have not been explored in previous studies. Obviously, in comparison with Design 1, Design 2 altered the interpretation of the substantial sources of error in the measurement and the G coefficients as well.

The residual component seems to be the largest source of variability due to the assessment conditions that occurred during test administration. This variability might be caused by noise in the corridor, psychological stress, tiresome, background knowledge students brought to the test, etc. In their study conducted on *"Using Generalizability Theory to Investigate the Reliability of Scores Assigned to Students in English Language Examination in Nigeria"*, seeking to test English proficiency, Akindahunsi and Afolabi (2021) interpreted the residual component along the following lines:

> The large residual variance captures both the person by item interaction and the random error (which we are unable to disentangle). Maybe some items were more easily answered

> by some participants or maybe there was systematic variation such as the physical environment where the test was administered, or possibly other random variation like fatigue during the assessment. Whatever the cases, these sources could not be disentangled from one another in this variance component. (p. 152)

From the above quote, it follows that the residual, as a source of variability in PT interaction, cannot be differentiated. It may refer to systematic or unsystematic variation caused by assessment conditions that were not controlled in the present measurement situation.

The second substantial variance component is related to **PH interaction** by 10.7% for relative error variance, and by 10.0% for absolute error variance. It was relatively low, because the students are homogeneous in terms of their vocabulary knowledge and writing ability, being exposed to the same language input and writng skills during terminal classes in secondary education.

Using one rater and four themes (tasks nested within themes) to measure the students' behavior, the depth of vocabulary knowledge test yielded a fairly high generalizability and dependability coefficients (0.93 and 0.92 respectively compared to Design 1 having 0.89). One justification for the high G coefficients obtained in Design 2 is the fact that the variability due to the raters and themes are both not substantial. Thus, it is not necessary to have several other raters and themes to measure the students' quality of vocabulary knowledge. Put differently, T:H variance component, the main effect for tasks nested within themes, accounted only for 0.6% indicating that the tasks contain themes of slight difference in the difficulty level. As to H variance, the main effect for themes accounted for 6.5% for the absolute error variance of the total variance, suggesting that there is some difference in theme difficulty among the four themes used in the current measurement situation. Overall, these values are negligible compared to the potential error variance due to the residual variance component (PT:H, e) discussed previously.

The generalizability of the overall vocabulary knowledge test scores was found to be fairly high (0.89/0.93/0.92) to measure the students' lexical ability. It confirms high degree of precision and reliability for the current measurement situation, especially

when the themes facet is inserted as a fixed facet. In consequence, the test in the current case was satisfactorily able to place the students relative to one another on the measurement scale. It was considerably able to differentiate among the students based on their own responses to the set of tasks nested within themes.

6.3. Relative Effects of Tasks, Raters and Themes on the Productive Depth of Vocabulary Knowledge Scores

This section aims to answer RQ3 associated with Design 3, the three-facet partially nested **PR(T:H*)*** design**,** which is penned as follows:

RQ3: What is the relative effect of raters, tasks and themes on the generalizability and dependability of the scores obtained from the vocabulary performance test?

Not surprisingly, both Design 1 and Design 3 analyses showed approximately similar results after the crossing facet interrelationship was shifted to partialy-nested, and the theme facet was inserted within Design3 (from P×T×R to P×R(T:H). The relative impact of the major variance components followed the same order, from the largest to the lowest, and had approximate degree of influence on the measurement precision. The G coefficients also remain the same (0.89) for both Designs. The results were relatively homogeneous eventhough the two designs estimate somewhat different sources of variability; seven variance components were calculated in Design 1 and ten in Design 3. This proximity refers to keeping similar facet observed levels for the tasks and raters, and nesting the tasks within the themes. The results obtained thus far are as follows:

1. In Design 1, the largest variance component refers to **PT interaction** by 49. 7% for relative decisions and 47.2% for absolute decisions. In Design 3, however, PT:H interaction was 42.4% for relative decisions and 40.3 % for absolute decisions, noticeably, it was slightly low than in Design 1.

2. In Design 1, **PR** was 29.8% for relative measurement and 28.3% for absolute measurement, but in Design 3 PR was 31.5 % for relative measurement and 30 % for absolute measurement. Both designs yielded PR component to be the second largest source of variability, and remarkably have slightly different values.

3. The **residual PTR,e** was 20.5% for relative measurement and 19.5% for absolute measurement in Design 1, and in Design 3 the PRT:H,e was 20.5% for relative measurement and 19.5% for absolute measurement. This means that both residuals refer to both systematic and unsystematic variation as the two designs occurred under similar assessment conditions (task relative difficulty, noise, participants psychological status, etc. which are unmeasured sources of error in the present research work).

4. In Design 1, the magnitude of error to the measurement precision was 0.230 for relative error variance and 0.236 for absolute error variance. In Design 3, the relative error variance was 0.233 and absolute error variance was 0.239, which are negligible values.

5. The third random-effects person × rater × (task: theme) analysis of variance estimated different variance components compared to the first person×task×rater crossed design. That is, no alternation happened at the level of rater facet (keeping 02 raters), and so for the number of tasks (08). Since all the raters scored the tasks nested within the themes (all the 04 themes and all the 08 tasks), and because all the students completed the tasks and hence themes, the G coefficient and the phi coefficients remained the same for the third design (0.89).

6. In Design 3, the residual or the interaction between students raters and tasks within themes (PRT:H, e) accounted for 19.5% together with all the eight variance components account for over 93% of the total relative error variance, at five. Additionally, for absolute measurement, there are eight contributors to the total error variance (PR by 30%, PT:H by 40.3%, PRT :H,e by 20.5% accounting for more than 89% of the total absolute error variance. The results obtained, here, are unique as never being investigated in previous research, in particular, integrating themes as a source of variability in vocabulary performance assessment via writing, using the context dependency measure proposed by Read (2000) and Read and Chappelle (2001).

Overall, the third partially-crossed three-faceted pr(t:h) design results indicated the relative effects of tasks, raters and themes on the examinees' vocabulary scores in the univariate G analyses. Of the tenth variance components, P, R, H, T:H, PR, PH, PT:H, RH, RT:H, PRH, PRT: H, the two largest contributors to the total variance are

the interaction between **students and tasks within themes** (PT: H by 42.4%), and the interaction between **students and raters** (PR by 31.5%). This indicates that variability in the students' performance across the four themes spread dichotomously along the tasks, and so for the mean scores, and that the students-by-raters effect was also considerable as the raters change their weighing across the individual students; the raters change the rating system as the student exam papers change. This has far being seen in the previous discussion of Design 1.

Unsurprisingly, it was found that the proportion of the theme-related variances contributing to the total variance was slightly underestimated in both designs (Design 2 and Design 3). The theme main effect was 6.5 % in Design 2 and 3.5% in Design 3. This means that there is some difference in the theme difficulty among the vocabulary tasks used in the current research. Subsequently, the participants performed differently throuought the tasks.

6.4. Impact of Number of Tasks and Raters on the Vocabulary Score Reliability

The aim of this section is to answer the D studies related questions and check the degree to which changing the number of observed levels can alternate the reliability of the test scores obtained in the optimization design phase. That is, whether the reduction of tasks and raters may lead to optimum reliability indexes. The questions thus posed are stated as under:

RQ4: What is the effect of decreasing the number of tasks designed to assess the vocabulary performance on the generalizability and dependability of the test scores?

RQ5: What is the effect of decreasing the number of raters on the generalizability and dependability of the test scores?

The objective of the present D studies is to inform assessment practitioners and stakeholders of the optimum number of tasks and raters necessary to achieve the best, or at least acceptable, estimates of generalizability and dependability coefficients, and the possible effect of error variance components on generalizability coefficients when the distribution of tasks and raters is altered. Findings from this study demonstrated that increasing the number of tasks and raters substantially affects the generalizability

estimates. Consequently, the generalizability estimates decrease as the number of the tasks increase, and the same does hold true for the raters. The task and rater distributions had a relative effect on the generalizability and dependability estimates.

The *what if analysis* indicated that decreasing the number of tasks and keeping the same number of raters increases the magnitude of error variances for both relative and absolute measurements. The same does not hold true for the generalizability and dependability coefficients that witnessed a decrease in the magnitude of the reliability index. This result is in consensus with Martinez's et al. (2007) study, which confirms that "the number of raters directly impacts measurement precision, or the confidence we can assign to observed scores as indicators of what students really know and can do" (p. 278). The researchers found that the standard error of measurement decreased when the number of raters increased and that adding a second (and may be third) rater results in a substantially more precise scores. They further suggest that making decisions about the number of raters is sensible and coast-efficient to join to the measurement situation.

Based on information obtained from the G studies, the effects of decreasing the numbers of the tasks and raters in a series of D studies were investigated. More precisely, once the G coefficient obtained in the G studies was 0.89, for cost effects the decision maker opted for reduction of the observed levels to optimize the measurement precision. The D studies indicated dissimilar patterns in both the relative and absolute G coefficients. For clarity, all the D studies with the relative G coefficients and phi coefficients are discussed in this section. The generalizability coefficients were compared descriptively to find out any differences in the generalizability coefficients occurred over the optimization design conditions or the D studies scenarios.

The aforementioned D studies showed that less tasks or raters are needed to reach an acceptable generalizability coefficient. To achieve generalizability of approximately 0.80 in evaluating the examinees' relative rank (0.82) and their absolute performance (0.81), if only one rater is applied, and about five tasks would be needed in the current vocabulary performance assessment. Besides, only four tasks and two raters would only lead to the relative generalizability coefficient of 0.83 and absolute generalizability coefficient of 0.82. Hence, to achieve a generalizability coefficient of 0.80 in the present

case, an application of either a combination of five tasks with one rater would be needed, or a combination of four tasks with two raters would be needed if we want to generalize the test results, and or use the test in future applications.

The data from this study suggests that, the second combination led to higher reliability than the first combination being under different conditions. Considering that both combinations are reliable, the decision maker may wish to opt for either depending on cost efficiency. However, to gain an overall vocabulary/lexical competency of third year secondary school learners, it is preferable to apply all the tasks with two raters combination, having led to 0.89 G coefficients for both relative and absolute interpretations. One more suggestion would be to divide the test into four tasks instead of eight and administer it on two different occasions to the same or different groups of learners.

It is important to mention that eventhough the approximations set for the optimization designs explained the obtained G coefficients in terms of Cardinet's et al. (2010) criterion value of reliability that states that, 0.80 is the standard value for judging scores as reliable. It is also possible to interpret these approximations and scenarios otherwise. That is to say, some researchers (e.g., Berk, 1979; Mitchell, 1979, as cited in Bottema-Beutel et al., 2014, p. 592) have determined coefficients of .70 and .60 as indicating acceptable reliability. These researchers assert that all G coefficients ranging between 0 and, with values closer to 1 indicate higher reliability and vice versa. Overall, the results of the D studies indicated that the reliability-like estimate differed somewhat substantially, depending on the number of raters or the number of tasks utilized in the D studies scenarios.

As previously stated in chapter three, the *what if analyses* procedure might be accomplished either by increasing or decreasing the facet observed levels. Of these, we opted for decreasing the number of observed levels, because the reliability coefficient was high. It was aimed at finding out the number of raters and tasks necessary to obtain an optimal reliability index and achieve cost-efficiency, one key quality criteria for designing, validating and applying effective performance/competence assessment. The research work is in consensus with the findings of Akindahunsi and Afolabi (2021),

who argued that "a decrease in the number of the items resulted in a decrease in both g- and phi coefficients in D-study" (p. 147).

So far, the reliability evidence has been considered across the research questions and has been arguably discussed via the quantitative data gathered. Validity, on the other hand, needs both statistical and non-statistical data analysis to be investigated. It rather needs supportive evidence that is why it is not discussed with reliability issues, but will be involved in the coming section devoted to general research results discussions.

6.5. General Research Findings Discussion

The results of the G studies and D studies displayed in this chapter lead to general conclusions. The findings emphasize multiple significant considerations for the use of performance assessments and application of G theory on the dependability and validity of the productive depth of vocabulary knowledge scores; the students' vocabulary performance across the tasks. The results thus produced would be discussed and accounted for along this.

6.5.1. Relative Effects of Tasks, Raters, and Themes on the Productive Depth of Vocabulary Knowledge Scores

Results of the G-studies in the three study designs have disclosed dissimilar patterns of difference in view of the percentages of the examinees, task, task within theme and rater-related variances contributing to the total variances (measurement error). **First,** in Design 1 we found that the largest source of vocabulary score variation was attributable to **person by tasks interaction** (person task variation) causing differences in the students' scores. Second, **students-by-rater variance** explained somewhat large portion of the total variance suggesting that raters were discriminant in assigning scores to students. Third, **the residual main effect** (students-by-task-by-rater interaction confounded with all unidentified sources of variance (e) demonstrated **relatively smaller** measurement error indicating that variances due to examinees' performance, task- sampling variability and raters variation and influences that are not considered in the study. Thus, PT interaction has an overestimated effect on the measurement precision.

Nevertheless, in Design 2, the largest contributor to the total vocabulary score variation was associated with the **three way interaction**, known as the residual component, which includes the variability due to the interaction between students, tasks within themes, confounded with random error and other unexplained systematic and unsystematic sources of error which were not involved in the design. The variance component of **PH interaction** explained a slightly smaller portion of the total variance (10.17%, 10.0% for relative and absolute error variances respectively). The variation in performance across the four themes was relatively low because the students are homogeneous in terms of their vocabulary knowledge and writing ability. In short, the eight tasks (and the themes henceforth) differed in the level of difficulty among themselves; except for theme 1 underpinning task 1 and task 2 where the task main effect was relatively low. The performance tasks were again differentially difficult for different students as illustrated in the person-by-task interaction component.

As to the **rater** facet, the G studies revealed that in Design 1 and Design 3 analyses the raters also contribute slightly to score variation in the vocabulary knowledge assessment conducted in the present research. Rater main effect was negligible (0.5 % in Design 1 and 0.2 % in Design 3) indicating inter-rater consistency.This finding is in agreement with Gebril's (2009) G study results, which indictaed that rater main effect had very closer values in the targedted task types (the reading-to-write tasks or integrated tasks accounted for 0.8% and in the independent tasks accounted for 0.00%).

In this study, leniency and severity factors could not be differentiated along the two raters because of low variation in the ratings obtained. This could be explained by the fact that the focus of the ratings is on vocabulary knowledge (meaning recognition, word formation and use) and not on writing proficiency/ability or writing quality. That is the raters are not required to assess the students' ideas, or deeply analyze the structure, rhetoric and organization of their performnace, this could shed some precision on the actual assessment especially across the context of raters models. This inter-rater consistency could also be explained, according to Gebril (2009), by the fact that the raters used similar scoring rubrics as a result of effective training sessions in the use of the scoring criteria that aimed to familiarize the raters with the scoring guide.

Consistency across the raters could also be interpreted in terms of the experience that these scorers have in scoring EFL essays.

Eventhough the ratings were consistent across raeters and tasks, the rater ain effect is considred a source of error in the current measurement. Previous research have indicated that rater facet is a dominant source of variancein the study of the impact of of teacher behavior on examinees' performance (Cronbach et al., 1972).

Students-by-rater related (pr) variation, on the other hand, explained the portion of 29.8% for relative measurement and 28.3% for absolute measurement in Design1 and 31.5% for relative error variance and 30.0 % for absolute error variance in Design 3. This is a relatively high proportion of PR contribution to the total of measurement error, which suggests that there was inconsistency in the rating severity and leniency for EFL students' performances as the rating schemes change whenever the students' paper chages. The students' language and style might be the direct cause of this inballanced rainings leading the raters to hold either positive or negative impressions on the performnces. This is a clear case of the impact that individual differences might have on the score variability. But when the raters' effect on vocabulary performance is compared to that of tasks and themes, it is as it sounds, small in size.

The percentage of the total variance accounted for by the variance in themes decreased from 6.5% in Design 2 to 3.5 % in Design 3. Actually, the variance that was attributed to the person by theme variance in Design 2 became part of the variance component of person by task and person by rater variance components in Design 3.

In Design 2, the value of H was 10% and in Design 3 was 5.1 % confirming that the theme facet is somewhat determinant of the students' performance; the students scored differently across the themes. This illustrates how far individual differences can have an impact on the dependability of scores. The students' schemata differd from one theme to the other, leading to variation in performance influencing their observed vocabulary scores.

6.5.2. Total Variance of Vocabulary Performance Scores and interaction components

The percentage of the total variance of the vocabulary performance scores is largely attributed to PT, and then to PR interaction components. Why is the percentage of the total variance of the vocabulary performance scores is largely attributed to the person by task interaction. In this section, we will provide two pervasive explanations for two pertinent actual sources of error; one is related to the tasks and the other is associated with the raters.

In Design 1, the first fully-two crossed design consistently indicated that the largest source of variance affecting score reliability (generalizability and dependability of test scores) was due to person-task interaction. one explanation may gain support from its overestimated impact on the generalizability of performance assessment with earlier work. The researchers like Shavelson et al. (1992), Shavelon and Baxter (1993), Lane et al. (1996), McBee and Barens (1998), Gao and Brennan (2001), Hébert, (2006), Liu et al. (2007), and Huang (2009) did find effects of person by task interaction on manipulating the generalizability and dependability of scores.

The person by task interaction effects has arguably proved to be high because the students' performance means scores vary from one task to the other and from one theme to the next; some learners scored well in some tasks and weak in others. PT interaction effect denotes that students use different strategies to solve the situation problems and interact within the given communicative situation. They change their strategies from task to task when trying out to construct their own responses. Given that the tasks entail different themes with relevant topical target words, the students were required to use the target words and use the language relevant to the target context. This linguistic performance may also differ from one authentic situation to the other to contextualize the target words corresponding to the written communicative contexts of Ancient Civilization, Education, Ethics in Business, and Feelings and Emotions.

Task variability has already been investigated by Parkes (2001) who confirmed that person-task variability is due to learners' inability to transfer their knowledge consistently from one task to the other. He thus associates this transfer with concept

mapping. Parkes considers error variance of person –by- task interaction "the most pervasive cause of poor reliability" and equates this source of error to problems of transfer of learning.

Following this line of thought, an examinee's misunderstanding of task context, instruction or target words makes him/her misinterpret task intention that would, in turn, deviate the student form his true performance of the task. Another reason for this source of variation, for Linn (1994), is explained by administering large number of tasks for performance assessment that substantially contributes to measurement error. One more justification for this source of error might be associated with the background knowledge students bring to the test (Shavelson & Webb, 1991). Learners' previous knowledge simply denotes their manage or mismanage towards the expected task performance. Having said, the students' performance might also vary due to the language used to do the task as the lexical ability is tested via writing and this variation contributes to the measurement error and score instability across the tasks. So, P×T variance component accounts for the differential performance of students across tasks; the students' scores vary from task to task.

Besides, task complexity/simplicity might also cause students to vary their performance along various tasks. Sasayama (2011), in his study on cognitive task difficulty and its effects on ESL students' written language production (task performance), he used two sets of picture-based narrative tasks: a simple writing task and a difficult writing task. He concluded that the difficult writing task elicited significantly more complex language production than did the simple writing task. Accordingly, those students' lexical performance changes throughout the tasks depending on the difficulty level and facility. Skehan (1998a) had already argued that individuals' attentional capacity is in fact restricted and sounds more difficult to pay simultaneous attention to complexity and accuracy when the cognitive load of tasks is heavier. An opposing view to Skehan's led by the interference model (Sanders, 1998) that stresses that task performance is not limited to learners' cognitive capacity but to the allocated time devoted to task completion as the latter factor may hinder processing sets of information. Simply because when learners are exposed to multiple stimuli

within a limited time, their area responsible for attention called the central executive working memory loses its control. Besides, learners' cognitive ability effect on performance, low performance can be reached in case of multiple complex tasks expected to be filled in in a limited amount of time. On the person task interaction effects, it is worthy to mention that, as Kamlasi and Nokas (2017) confirm, in writing learners tend to integrate a set of skills and knowledge as it is described as more a complex process, this procedural complexity made it unexpected that similar learners will perform equally well on different tasks in EFL writing performance. The same can be said for lexical performance as any language test is considred a vocabulary test

Previous studies have indicated that variability due to PT interaction presents an issue that performance-based assessments often encounter (Shavelson et al., 1992; Shavelson et al., 1993; Lane et al., 1996; McBee & Barens, 1998; Gao & Brennan, 2001; Hébert, 2006; Liu et al., 2007; Huang, 2009). PT variance component arguably proved to be a drawback for performance-based assessments (Shavelson, Ruiz-Primo & Wiley, 1996; Hébert, 2006), because "task-related variance causes scores from performance assessments not to be generalizable and thus inappropriate for high stakes use" (Parkes et al., 2000, p. 397). The two variance components of task main effect and person-by-task interaction variance relatively affected the generalizability of the present in depth vocabulary test scores.

As such performance based assessment is questioned for its reliability of scores (Miller & Linn, 2000). This is mainly because it is based on multiple constructed response open-ended performance tasks, where learners' performance varied unlike MCT containing predetermined options were examinees have to select right responses. Constructed response formats do not only increase sources of error at levels of task variability (person task interaction variance component) but also at level of rater judgments.

Amongst the research findings is the person rater interaction considerable effect on the accuracy of the vocabulary test scores. One logical justification for the high effect of this source of error might refer to rater drift. Performance assessments, as previously mentioned in the literature, require multiple judgments to yield accurate results.

Neverthless, rater drift might affect assessment scores as raters might change their scoring behavior not only across tasks but also across individuals. Very long ago, Thurstone (1927) highlighted the role of individual differences in judgment and perception. Originally, Wright and Douglas (1986), the leading figures in performance assessment and constructed response formats, assert that raters in charge of grading a particular constructed response performance should be free from subjectivity and similar scores ought to be assigned. Contrarily, rater drift exist no matter how raters try to control and guide their perceptions and strictly transmit their judgments. For them, rater drift involves changing scorers behavior across different test administrations or even across similar test administration, and this can influence universe scores.

In a nutshell, besides the aforementioned arguments for score variability and sources of measurement error in assessing vocabulary via writing, Schoonen (2005) mentions other several sources responsible for error variance affecting the writing performance. The topic students required to write about, the discourse mode, text type or genre (description, exposition, narrative or argumentation), the time allocated to do the task, the writing mode (paper-and-pencil or text processor), the testing conditions (e.g. psychological status, noise, light), rater inconsistency, scoring procedure whether holistic or analytic, and traits to be scored, namely content, language use or spelling.

It is worth mentioning that there exist several points worthy to highlight in the discussion of the research findings particularly related to the assessment precision. These are stated along the following lines:

1- We noticed no variance in G coefficients obtained from the G studies. The coefficients were (0.89) for both relative and absolute measurements, because the results of both relative and absolute error variances are slightly different across the sources of variance; across the tasks and raters. The G coefficients demonstrated how the examinees' scores are generalizable and dependable. They led to a satisfactory measurement precision across the test items in the overall assessment of vocabulary knowledge whether these information used to rank-order students for relative decisions or to compare their performance against the expected level of performance (absolute

decision). These values extended a coefficient of .80 a conventional value that has been taken as a standard value to interpret scores reliable (Cardinet et al., 2010).

The coefficients were fairly high because of the heterogeneous population sampled from the universe of admissible observations, as explained by Cardinet et al. (2010) based on Cronbach's conception. Cronbach cautioned against probable erroneous interpretations. The sampling procedure of the objects of measurement (students in our case) can dramatically have an impact the value of G coefficients. The more heterogeneous the target population, the higher Coef_G will be. The converse is equally true. The more homogeneous the population, the more difficult it will be to differentiate between its members. It could be the case here.

2- The values for the standard error of measurement were negligible for both relative and absolute measurements.

On a 0–1 scale, the SE was 0.230 for the relative SEM and SE: 0.236 for the absolute in Design 1. In Design 2, the total error variances for both the relative and absolute were SE: 0.207 and SE: 0.215 respectively, and in Design 3 the relative SEM was 0.233 and the absolute SEM was 0.239. Eventhough measurement error is a negative sign of inaccuracy, it can differentiate between a measured score and its true score. The values obtained for the SEM proves high degree of precision and reliability for the current measurement situation besides the G coefficients that were very closer to each other (0.93 for Coef_G Relative measurement and 0.92 for Coef_G absolute measurement obtained in Design 2 and 0.89 G and phi coefficients in Design 1 and Design 3.

3- Treating themes as fixed in the fixed effects model had a substantial positive effect on the dependability of the productive vocabulary knowledge test, yielding very high G coefficients

In Design 2, the two-facet partially nested P(T:H) design with H considered fixed, a fairly high generalizability and dependability coefficients (0.93 and 0.92) were obtained. These results support those findings obtained by Akindahunsi and Afolabi (2021), where the G coefficient was 0.90 and Φ coefficient was 0.87, those findings of which is an indication of high reliability of scores. In a mixed model G analysis with

scales considered fixed, Tobar et al. (1999) found generalizability to be quite high (0.96). They indicated that altering the assumption from random sampling to fixed sampling would have a dramatic impact on the G coefficients that is, in our case, by assuming that the themes are a fixed facet, the G coefficient increased accounting for G= 0.93 and phi= .92 more than it would if themes were considered random.

Considering the facet of themes fixed means that we assume that we have no intention in generalizing beyond the four themes upon which we contextualized the tasks prompts. Because the fixed facet for themes involves the whole universe of themes, and because the four themes define the construct of interest, there is no reason to generalize to the universe of possible themes.

As it has already been discussed in the literature, when fixing a facet, the resulting G coefficients will be high, because fixing a facet increases score variance and reduces measurement error resulting in an increased estimated reliability (Meyer, 2010; Cardinet et al., 2010). The relative and absolute error variances were reduced in this study due to fixing the theme facet. The SE was 0.230 for the relative SEM and SE was 0.236 for the absolute in Design 1, in Design 2, however, the total error variances for both the relative and absolute decreased to SE: 0.207 and SE: 0.215 respectively.

The results demonstrated that the DPVK test scores were highly dependable. This entails that the students' scores were highly dependable in terms of reflecting the lexical ability of the applicants. One argument for this, as stated in chapter one, is that fixing a facet increases the G coefficient, but "precludes inferences beyond the conditions of the facet included in the measurements" (Tobar et al., 1999, p.148). Still, the generalizability analyses resulted evidence of the validity of the student performance over the targeted themes in this study. These sizable coefficients could be explained in Brennan's (1992) terms as the variance attributed to the theme facet is absorbed into other variance components by taking the average over the conditions, or levels, of this H fixed facet.

6.6. Collecting Evidence for Validity Using Messick's Unified Validity Framework

RQ6: What is the relative effect of tasks, raters, and themes on the construct validity of the vocabulary performance assessment?

Above all, the purpose of this section is to show the applicability of G theory to address validity issues besides its utility in assessing the dependability of the observed scores discussed so far. Once the title of this thesis is *"An Investigation of the Reliability and Validity of Vocabulary Performance Assessment Using G Theory"*, the second focus of this research work is to examine the validity evidence that actually exists to support the uses of the current in-depth vocabulary performance test, and so for the rubrics used to obtain the test observed sores. It is worth mentioning that issues of reliability evidence had been already discussed in the previous sections of this chapter. Unlike reliability, validity is not purely quantitative in nature but it is evidence-based. In this part, we use data analysis and psychometrics to ensure the validity of the current assessment outcomes and score interpretations.

G theory "is well-suited to investigations of construct related-evidence of validity because a single generalizability investigation may provide multiple inferences of validity" (Kraiger & Teachout, 1990, p. 19). Being the principled approach in this study, it enabled us to estimate the testing conditions, the multiple variance components, simultaneously. Using G theory, it was also possible to determine the relative effects of variability attributable to the measurement facets: tasks, raters, themes and interactions among these facets and the object of the measurement and their effects on the estimation of the students' PDVK together with systematic and unsystematic errors. More importantly, it informed us with how these facets and their interactions are more or less responsible for the variability of test scores. These data will help us interpret the results and draw inferences especially related to addressing validity evidence.

It seems significant to specify aspects of construct validity as types of evidence that would be required for various types of uses and interpretations of the results of this in-depth vocabulary knowledge test. Aspects of construct validity collected to support validity evidence are neither exhaustive nor exclusive, rather categorized under four sub-aspects of construct validity: (1) generalizability, (2) convergent validity, (3) content analyses (content validity), and (4) structural (internal) validity.

6.6.1. Generalizability

Perusing Messick's evidence-based argument framework, generalizability is one evidence supportive of validity. The findings indicated that the PDVKT produced dependable scores all through the three G study designs. The reliability index was indicated by the size of the G coefficients (0.89; 0.92; 0.93) obtained across the three G study designs and D studies where the G coefficients were maximized under a set of conditions. Generalizability as reliability (Crocker & Algina, 2008), an aspect of construct validity, is explained in terms of the consistency of the students' vocabulary performance across the tasks, raters and themes of the actual assessment. The scores are consistent in the universe of generalization, which is limited in scope.

G theory was used to collect evidence of construct validity. To this end, we need to provide evidence that the test results are dependable across the tasks (choice of context, instructions, target words), across the raters (scoring procedure) and across the themes (four diversified topics). Within the three G studies, we conducted separate G study designs: 1) between-task generalizability, 2) inter-rater agreement, and (3) generalizability across themes (contexts).

1) *Generalizability across the tasks:* the variance across the eight tasks was relatively low, but the person-by-task interaction effect was substantially high, but this does not mean that the scores are not consistent across the tasks. Because previous research proposed that in performance-based assessment there exist substantial task-specificity which means that two tasks seeking to assess similar knowledge and abilities (problem solving skills) frequently result in sizeable differences in performance (Linn, 1994).

2) *Generalizability across the raters:* the variance attributable to the ratings was low, thus the scores are consistent across the raters.

3) *Generalizability across the themes:* the variability due to the themes diversity is also low, which indicates that the students scored consistently, even differently across the different themes **(**T:H= 0.6 which means that the inter-theme variability is low indicating low inter-task variability).

Overall, the scores obtained were consistent across the three research major facets and thus the level of generalizability arrived at is acceptable; this shows that the test is valid. Generalizability, in our context, is taken to mean validity in Messick's terms. This fact extends the generalizability or external validity inference of the ratings, and of the students' behavior across task performance.

6.6.2. Convergent Validity Evidence

Basically, when investigating the issue of construct validity, the researcher ought to find out what sources of variability contribute to test score variance (test performance), what components that constitute error, and what is the relative size of variance components (Cronbach & Meehl, 1955; Kraiger & Teachout, 1990). To illustrate this application, we analyzed the results obtained from the G studies to investigate the variance components that are interpretive of construct-related validity evidence (convergent validity) of the actual measurement. The expected magnitude of variance components is explained in relation to the total variance and number of facets included in the design.

Results from the G Studies Supportive of Construct Validity: convergent validity is evident in G theory settings when scores show invariance over facet conditions (Kane, 1982). In the present measurement situation, by facet conditions is meant multiple tasks, multiple raters and multiple themes. In other words, convergent validity is evident when students are "similarly ranked over methods" (Kraiger & Teachout, 1995, p.4). i.e., if students are similarly ranked over rating forms, or tasks, or themes, or methods, etc. As such, convergent validity should be interpreted in terms of variance in students across facets of measurement (tasks, raters and themes) and the magnitude of variance in relation to the total variance and number of facets included in the design. The findings supportive of convergent validity evidence are summarized in Table 6.1:

Variance components	Expected magnitude	Convergent validity evidence
T	Small	- Scores are invariant over PDVT tasks. There existed some difference (minimal) in task difficulty (what made students produce approximate scores across tasks?)
R	Small	- Scores are invariant over rating forms since neither rater severity nor leniency are marked.
TR	Small	- Rating differences across tasks are slight (interrater consistency)
H	Small	- Scores are invariant over PDVT themes. The main effect for themes on students' performance is low.
T:H	Small	- Inter theme variability is low indicating low inter-task variability. Scores are invariant over PDVT themes. There existed some difference (minimal) in theme difficulty (what made students produce approximate scores across themes?)

Table 6.1: Expected Results Supportive of Convergent Validity Evidence

The estimated variance components appeared in the table above have low magnitude of variance. They displayed invariance over the assessment facets. They therefore support convergent validity evidence for the PDVKT. Subsequently, high convergence over the facets of tasks, raters and themes have been reached. In the coming section, content validity is also used as an evidence of validity and as argument of why the test has achieved a certain level of convergence validity and generalizability.

6.6.3. Content analyses/validity

In the process of test design, development and implementation, there are several steps that can be regarded as a source of validity evidence for the test itself and its intended use. The various methods used in the design of the measurement procedure permit for the accumulation of validity evidence (Hill et al., 2022).

To ensure construct validity, and content validity in particular, various consecutive methodological procedures were applied in the construction and validation of the vocabulary performance test. See Chapter four for a full description. First, before writing the first draft for the test items, the purpose of the test and the construct to be assessed were identified together with the content standards and test specifications.

Once the defined content standards were established with reference to the textbook "*New Prospects"* units'/language outcomes, the current assessment was intended to be compatible with those standards as "the first validity question concerns the adequacy of the alignment of the assessment with the content standards" (Linn, 1994, p. 568). To judge the pertinence of alignment of the assessment with the content standards, a board of subject matter experts were asked and the test content was designed accordingly. The textbook that served as the beginning of venture provided the blueprints for the assessment; even though the performance standards are implicitly stated in the units outcomes. The test specifications were written despite their implicit status in the syllabus.

The test items were then reviewed by content area experts who provided feedback to ensure content validity evidence. The test content was refined and revisited accordingly. These experts were required to judge content relevance to the construct of interest. See Appendix C for the remaining quality criteria used to review the test items by experts to ensure whether the test measures what it is supposed to measure, or if the tasks are representative enough of the construct under study. After refinement of the test content, the test was piloted to a small group of respondents to further gather evidence of content validity. To check how the respondents actually engaged with the imminent test, they were tasked to do the test and put comments after each task describing their difficulties concerning the target words, the tasks prompts and instructions, etc. They were also asked to fill in a questionnaire gathering qualitative and quantitative data concerning the quality of the test and its items (e.g. time allotment, alignment with the course content, clarity… etc.). Furthermore, before administering the test, reliability analyses was conducted using G theory and ANOVA procedures to support the validity evidence of the piloted test.

Data from these methodological procedures are used to support validity evidence for the actual test, as the procedures of validation were taken as firmly as possible. However, this does not mean that the validity argument had been collected rigorously. The test should be practiced in new contexts for different universes of generalization to guarantee its wider generalizability.

- The largest variance components affecting the convergent validity of the students' performance scores throughout the three research G study designs:

The greatest threat to our test validity is attributable to PT effects that proved to have a considerable effect on the score generalizability. The research findings revealed that the person-by-task interaction variance affected the construct validity (convergent validity in particular) of the students' observed scores. This explains the students' shift in their use of strategies throughout the tasks completion; their performance varied because they used different methods across the different situations. This further illustrates why they scored differently from one task to the next; they highly scored in some tasks and less in others. Therefore, the convergent validity evidence is relatively low. The score variance is attributable to individual differences or to task-to-task variation in the performance of individual students. PT interaction effects reflect within-task instability in the examinee's individual performance. This effect demonstrates that the relative ranking of the examinees varies considerably across different tasks. In sum, the students are differentially ranked by PDVT tasks; the students vary in task experience and abilities to perform the various tasks. It is clear that PT variance component did not indicate invariance across tasks.

The second threat to convergent validity was PR interaction effect which has a considerable variance effect on the ratees' performance. Low convergence over ratings means that individuals are differentially ranked by the rating scales; when forming judgments, the raters were roughly dissimilar in how they scored the paragraph responses across the students, the ratings or rubric implementation might be changed from one individual student to the next.

6.6.4. Internal Validity

In the previous chapter of data analysis, we introduced a statistical procedure used to assess the internal structure of our assessment. This procedure intended to justify the validity of our test in Design 2 and in Design 3 after inserting the theme facet into the G analyses. Item facility as a procdedure revealed that the test could reliably differentiate between the tasks and hence the tasks netsed within the themes in terms of their difficulty and facility, thus enabling us to rank order the tasks and themes. The

order is as follows: H1 (T1 then T2), H4 (T7), H3 (T6), H2 (T4, T3), H3 (T5) and H4 (T8) with tasks 5 and 3 are of equal difficulty. According to Brown et al. (2019), this mathematical mode confirms that not necessarily each item of a scale (in our case each theme or each task along the four themes or along the eight tasks respectively) has the same value or duplicates the same difficulty level. This test has internal validity as it could identify both the students' lexical ability and item difficulty of both themes and tasks, generating a hierarchical scale of the items from easy to difficult.

Up to this point, the psychometric quality of the present test have been examined using G theory. Reliability and validity issues were discussed based on the estimated variance components and G coefficients produced in G studies. This study therefore calls for a number of recommendations for assessing the assessment applying G theory that are presented later in this chapter.

6.7. Implications and Caveat

From the aforementioned discussions, we come to the conclusion that the current test is not perfectly accurate since the scores and ratings obtained are subject to measurement error like any other tests or measuring procedures such as questionnaires. One reason for that, is that the construct and, even the conditions underlying its measurement, that we are seeking to estimate are difficult to define in an absolute manner, and is not bound to direct observation (for productive depth of vocabulary knowledge cannot be directly observed). But above all, we attempted to develop a research instrument that we assume is capable of eliciting evidence of the trait under study. Inevitably, however, various conditions, such as individual differences, task variability, inter-task consistency, raters and inter-raters consistency, theme variability and the interactional effect relationships existing between these conditions have a relative impact on this process of measurement and yield variability, and hence produce errors in the data set or findings. To cater for these challenges, we tried to quantify as many as possible variables affecting the measuring procedure and scores obtained looking forward to control them to achieve optimal measurement precision.

To this end, a methodology based on G theory framework was essentially applied in order to provide an estimation precision of the actual measurement situation's models

and designs; being bound to multiple variance components. This approach made it ultimately possible to estimate and depict the extent to which the current test is dependable besides that, it is simultaneously informative of the different sources of error and their contributions to the measurement precision and to the total standard error of measurement. Depending on the magnitude of sources of variance and their contributions to the measurement error, we could invest these variances to improve the test, the measuring procedure, by means of optimization designs seeking to obtain maximum reliability for further implementations. A measuring instrument having gone through optimization phases can for instance, serve as a frame of reference for entrance or placement tests purposes being able to design and evaluate measuring instruments.

Using G theory principles, the current study indicated that the reliability was high, which affirms that the scores assigned to third year EFL learners were dependable and generalizable. These findings provide a preliminary psychometric evidence to suggest that the vocabulary knowledge test has academic utility and may be a valuable tool to assess and collect important information about students' level of competence in vocabulary. The latter may serve to convey as prerequisite knowledge to students before being inducted to university level, and thus determine secondary school EFL learners' exit profile they bring to peruse their higher education.

In this sense, this test is suggested to be applied to examine the new BAC holders' level, quality of vocabulary knowledge, and competence in the use of previously acquired vocabulary to place learners at appropriate groupings. One more suggestion is to develop more test modes to test productive and depth of lexical knowldege and to apply other psychometric theories such as Item Response Theory to validate such a kind of tests to be adopted as standerdized exams or entrance tests used for higher education placement purposes.

In general, the results of this study make an appeal to assessment stakeholders towards an achievement of measurement precision; using G theory a researcher can investigate whether a given assessment is reliable/dependable (consistent) and valid (accurate). The two indices of generalizability- relative and absolute G coefficients- provide an index of reliability for any measurement procedure and accounts for

measurement precision. This theory is more informative of the magnitude and different sources of measurement error and their contributions to the total measurement error and their effects on measurement dependability. Besides, within its framework, G theory application offers an optimization design that allows assessors to reduce or increase the facets levels to reduce the amount of error to achieve optimal level of reliability.

The findings of this research partly suggest that different facets affect EFL vocabulary performance. The pedagogical implication entails that different sources of variance are involved in vocabulary assessment via writing (comprehensive-embedded test); a fact that teachers of English, test developers, and researchers should be aware of their undesirable effects for test validity and reliability. Based on the study results, we can safely assume that the variance components of tasks, raters, and themes, among other possible facets be it occasion, are key facets affecting the generalizability and dependability of vocabulary performance scores. Interaction among the defined study facets (interaction effects of students, tasks, rater, theme, and the residual component) are confidently regarded as measurement errors. Bachman (1990) considers interaction between student, rater, tasks, and other facets, sources of variability in any assessment.

Because the students' performance (students-by-tasks interaction) proved to be the largest variance component contributing to the total variance, the validity and generalizability of a measuring instrument could be featured through an assessment of depth of vocabulary knowledge construct with different tasks, raters and themes. Considering vocabulary component part of assessing writing proficiency, we would adopt Marcoulides' (1989) suggestion that validity could be ensured by means of assessing writing proficiency with different scoring methods, different raters and tasks. Test convergent and discriminant validity could also be investigated through the generalizability analysis (Schoonen, 2012).

When conducting D studies, G theory allows practitioners to investigate the major variance components in productive depth of vocabulary knowledge/performance assessment, and optimize reliability and generalizability indexes of measurement in this content area and among other areas of EFL assessment. Optimization designs that could elicit students' true scores (true vocabulary abilities/productive knowledge and use)

could be achieved through decreasing the number of facet levels "with spending the minimum cost, energy, and time brings the fairness to measurements in an educational setting by providing an in-depth analysis of factors, which affect the observed scores". (Khodi, 2021, p. 23)

Conclusion

In this chapter, we discussed and interpreted the research findings in an attempt to answer the six research questions together with the posed sub-questions. We examined the psychometric properties of the data obtained from the vocabulary test by estimating the generalizability (g) coefficient, phi (Φ) coefficient and construct validity (content validity, generalizability, convergent validity, and internal validity). In doing so, we have provided an interpretation to the percentile values obtained from the G analyses and optimization procedures. In general, the test can be said to be a dependable measure of testing vocabulary knowledge of first year degree students or new baccalaureate holders exposed to *"New Prospects"* textbook content, since it displayed high G coefficients. Actually, given one perspective of the present research, most of the tasks and thus themes could be reduced to four tasks and hence to two themes respectively yet still retaining good psychometric properties and yet reaching an acceptable reliability index.

What is worthy to consider is the fact that even the current test has arguably proved to be reliable and valid, it is still prone to measurement error. Accordingly, this chapter stresses the need to embark on studies to examine other possible sources of variability in vocabulary performance assessment such as, occasion, teaching method, rating scales, and to implement other theorires of psychometrics to examine the dependability of this construct. The chapter, finally, ends with a number of recommendations that are suggested to assessment stakholders to consider G theory in quest of validating and establishing standardized assessments.

GENERAL CONCLUSION

This research aimed to examine the reliability and construct validity of performance assessment scores obtained from an in-depth productive vocabulary knowledge (DPVK) test that targeted first year EFL ENSB students' ability to recognize word meanings, employ appropriate word formation processes, and use target words in appropriate communicative contexts. It is unquestionable that productive knowledge of vocabulary is highly critical to their academic achievement at the university level. However, due to the inconsistencies of test scores they obtain especially from open-ended questions, which are believed to tap performance levels, EFL students often question equity and transparency of assessment procedures and so for the applied scoring systems. For some students, scores are suspect and do not really represent their true ability in the language. The research, therefore, was an attempt to reveal the multiple sources of error affecting the current assessment precision and estimate the magnitude of error threatening the consistency (reliability) and accuracy (validity) of performance/competency scores by means of investigating the relative contribution of three facets of tasks, raters, and themes to the depth of productive vocabulary knowledge (DPVK) assessment precision.

To complete the investigation, a review of literature and a field investigation were necessary. These appear in the two major parts that make up the body of this thesis. The first part of the thesis has provided a literature survey on G theory, vocabulary assessment and performnce-based assessment that resulted in a set of concepts and principles, which mostly represent the study frameworks (statistical, lexical and validational). In doing so, G theory conception, evolution and merits were discussed in chapter one in connection with the shortcomings of CTT in a fairly exhaustive thorough manner to deepen understanding of its core concepts and principles. The stages of G and D studies together with the different study designs and models were also given concern and were examined thoroughly, with a shifted focus to reviewing previous research on the theme under study in addition to how well the theory contributed to the research endeavor.

Moving from general to specific, the main issues that chapter two has addressed were related to the significance of vocabulary knowledge, its conceptualization, and vocabulary depth delineation. Then, it explored various issues related to the assessment of vocabulary knowledge (and vocabulary depth henceforth) such as the type and nature of vocabulary assessment procedures, approaches, and methods. The study has revealed a number of difficulties associated with vocabulary validity and modelling its assessment, upon which an attempt was made to reach a conceptual framework that might allow to design and develop DPVK test for the assessment of lexical competence. The framework that was elaborated within task-based performance paradigm, emphasized test congruence with the purpose and context of the inquiry (stress on context dependence vocabulary measure).

On a narrower scope, the genesis of performnce-based assessment and its definition were examined in chapter three, and its characteristics were examined on a broader scope informed by scholarly works. Shifting interest from objective to more performnce/competency assessment, the key quality criteria, including reliability and Messick's construct validity, were emphasized as they serve as foundational grounds workable in designing performnace tasks and in the validation process of DPVK test. Additionally, an examination of the relevant scoring rubrics and factors affecting performnce assessment psychometric properties were also stressed in the chapter.

In the second part of the thesis, a methodology based on the multifaceted approach of G theory framework was essentially implemented in order to provide an estimation precision of the actual measurement situation was fully described in chapter four. Within the scope of this current exploratory study, G theory, a powerful statistical method, was used to gain insights into the potential of applying the psychometric measurement conditions and theoretical concepts of G theory to dig up the difficulties that the sampling variability of performance/competence-based assessment might yield in. In other words, the theory was utilized to gain initial insights into the major variance components attributable to the measurement error of the present measurement situation upon which the D studies would build up their approximations in a "What if Analysis" procedure.

This research was partially descriptive, the aim of which was to provide a comprehensive and accurate image of the variance components relative impact on score variability by collecting data quantitatively. The quantitative data collection procedure undertaken was a test of DPVK which has been developed according to Nation's (2001; 2013) framework of what is involved in knowing a word, and within the communicative task-based endeavor. Accordingly, eight communicative tasks were administered to tap into EFL students' lexical competence, and their actual performance was scored by two experienced raters via holistic rubrics made up of three basic distinctive constituents including word meaning, word formation, and word use that comprise learners' whole targeted lexical competence. These data provided measurable scores for the psychological attribute of students' vocabulary mastery, which could indicate whether their receptive vocabulary knowledge becomes productive. Bearing in mind that the optimal goal of this study was assessing the assessment, we emphasized the psychometric characteristics of the test itself irrespective of what it was planned to measure. Put otherwise, the construct being assessed presented a source for data collection to provide reliability and validity evidence and, thus, indicate if the inferences and interpretations of test scores are trustworthy and generalizable.

In order to estimate sources of error that might have affected the consistency and accuracy of the test scores in this psychometric research, three G study designs followed by four decision studies were carried out to analyze data systematically to further explore the reliability and construct validity of the performance assessment. These procedures have been described in chapter five in two different stages. In the G study stage, two-facet-fully crossed p×t×r, two-facet partially nested P (t:h), and three-facet partially nested Pr(t:h) designs were conducted, where tasks, raters and themes (context) were considered sources of error variance, with students as objects of measurement and facets were treated as random in the first random crossed facet design and fixed in the two latter mixed facet designs. Using the *EduG* and *ANOVA* procedures, it was possible to compute and analyze data, and hence partition the magnitude of measurement error into percentages of variance components attributable to tasks, raters and themes.

In the D study stage, the objective revolved around an investigation of how many tasks and raters are needed to achieve a maximum of reliability index. The what if analyses was concerned with three optimization designs: decreasing the number of tasks, decreasing the number of raters and finally decreasing the number of tasks and raters simultaneously.

The last chapter has displayed the quantitative results obtained from the vocabulary test concerning both G studies and D studies. The G studies analyses displayed that the generalizability of the DPVK test scores was found to be fairly high (0.89) to measure students' lexical ability. It, thus, arguably proved a high degree of precision and reliability for the current measurement situation. In effect, the test was satisfactorily able to place students relative to one another on the scale of measurement as it could differentiate among the students based on their own responses to a set of tasks nested within themes (tasks being embedded within themes).

Although demonstrated high level of generalizability and dependability to test first year EFL university students, as an entrance/attainment test, or to test the baccalaureate students, as a summative test, our test was affected by various sources of error. In fact, the G studies revealed that the largest sources of variances affecting the generalizability and dependability of the students' vocabulary competency scores were attributed to students-by-task interaction, students-by-rater interaction, and the residual components (**PTR,e** and **pt:h,e),** which stands for the variability due to the interaction between students, tasks and raters confounded with other unexplained systematic and unsystematic sources of error; and the interaction between students and tasks within themes confounded with other unexplained systematic and unsystematic sources of error respectively.

The G studies findings also showed variances in the generalizability and dependability coefficients of students' competency assessment scores in solving the communicative tasks obtained across the three study designs. These were explained by the nature of facet level sampling, whether fixed or random. When the theme facet was considered random, the G coefficient was 0.89 for relative and absolute measurements

and when treated as fixed the G coefficient reached 0.93 for Coef_G relative measurement and 0.92 for Coef_G absolute measurement.

Further findings regarding the construct validity of the test revealed that, the largest variance components threatening the convergent validity of the students' performance scores throughout the three research G study designs were attributable to the student-by-task interaction variance and student by theme interaction, which illustrate students' shift in their use of strategies throughout the tasks/themes completion; their performance varied because they use different methods across the different situations (scored differently across tasks and hence across themes).

Additionally, the second threat to the test consequential validity was the PR interaction effect which had a significant variance effect on the students' performance. Low convergence over ratings illustrates that test takers were differentially ranked by the rating scales; raters roughly judged the responses and they practiced the scoring rubrics dissimilarly across the students.

Another major finding was concerned with the D studies approximations which found that the generalizability coefficients were different across study designs for both the reliability index and validity evidence of the test scores. Besides, decreasing the number of tasks and raters decreased the generalizability coefficients, and in the meanwhile, increased the magnitude of error variances for both relative norm-referenced and absolute criterion-referenced measurements. Research results also revealed that only five tasks with one rater or four tasks with two raters were required to obtain acceptable levels of generalizability, thus optimizing the measurement precision. On the basis of the above findings, a number of recommendations were proposed so that assessment stakeholders can apply to improve vocabulary performance/competency assessments.

Since any test is prone to measurement errors and evaluating students' vocabulary learning is at the core of the teaching learning processes, assessing the test itself is generally "questionable" at the level of language test development and validation. Grades drawn from summative or formative assessments are to be transmitted to students who might question their fairness and transparency. Decision making such as

passing a year is based fundamentally on the test results, it is therefore, necessary to ensure how far a test is fair, precise and above all valid. Reliability and validity seem to be unstable assessment factors in performance/competency assessment paradigm. The goal of this psychometric study was to improve assessment and hence better educational decision-making.

The pedagogical implication entails that teachers of English, test developers, and researchers should know about the undesirable effects of test validity and reliability. The various components such as tasks, raters and themes are inevitably involved in vocabulary assessment via writing; in a comprehensive-embedded context dependent measure of vocabulary, especially when constructed within task-based performnce. More importantly, using G theory assessment practitioners can estimate the major sources of variability in productive depth of vocabulary knowledge and performance assessments conducting G studies, as they can improve reliability and generalizability indexes of measurements via D studies.

In general, in line with the study findings, it is recommended to use G theory to address issues of assessment precision. Incorporating a statistical method, G theory allows assessment stakeholders to investigate whether a given assessment is reliable, dependable, generalizable and valid. It is possible to count the magnitude of sources of measurement error and their contributions to the total measurement error and their impacts on measurement dependability. It also allows for reduction of measurement errors seeking to arrive at optimal levels of reliability index. In sum, we are of the conviction that taking these recommendations into consideration in future measurements will help in validating and optimizing levels of reliability, especially for developing standardized vocabulary assessments.

Limitations of the Study

Overall, the study has some limitations, among which the halo effect that might affect the universe score variance given that the students' performance was rated by secondary school teachers, whom we assume will tend to have positive impressions of the Baccalaureate holders being graduated from their schools. Therefore, there is a need for more secondary school raters judging the performance or rather rating it from the

perspective of teachers belonging to other sectors, be it university or middle school teachers, or inspectors, or even researchers to prevent bias. Furthermore, the students' performance ought to be scored by more than three raters to obtain more accurate and reliable results.

One more limitation associated with rater drift concerns the two raters involved in the assessment. They had different levels of experience that might produce inconsistency in the observed scores. Although both received enough and equal training and followed the same rating protocol, this discrepancy in professional experience may pose a potential risk to inter-rater reliability.

Occasion, as source of measurement variability can affect the measurement precision, and thus ought to be another facet investigated. Because it was very exhausting for leaners to respond to eight tasks in one test occasion. In order to save students' time learning, it was only possible to take two hours session rather than more sessions presenting the teaching load and time allotment. Linn (1994) asserts that it is impractical to increase the number of tasks in performance-based assessment as the fact largely contributes to measurement error.

The intention of the study was to generalize on a sample size of themes selected from the defined textbook. This implies that no intention was made to generalize beyond these four themes to others and it would not be reasonable to do so. More possible themes can be observed levels in this study. It is suggested to select words and themes from a variety of world wide word lists such as the British National Council and Academic Word List, besides textbook vocabulary addressed to teaching learning purposes.

Finally, we need to admit that the small size is one limitation of the study. Eventhough, it proved to be representative of the entire population because of its sampling procedure it did not tackle all the possible EFL students belonging to different universities as it focused only on ENSB students, the universe of generalization is limited to first year ENSB students enrolled in the Department of English.

Suggestions for Further Research

On the basis of the research findings, research orientations, and the research limitations it is possible to suggest recommendations for further research proposed in quest of enriching the educational research literature. These are:

- It could be suggested that further studies need to be conducted to investigate the halo effect. Without forgetting the accuracy and consistency of the rating scales used in the study, because no matter how they are interpreted by the scorers, they are affected by assessors' subjectivity when assessing performance. Furthermore, it is suggested to investigate the holistic and analytic rubrics and their relative impact on test score dependability and conduct contrastive analyses for the results gained by each using G theory principles.

- It is important to explore the sampling variability due to occasion (test-retest reliability) as a source of variance affecting the measurement situation precision specifically in terms of the reliability of obtained scores.

- To achieve a high level of generalizability for low or high stakes assessments of vocabulary or any other domain of testing within the performance/competence assessment framework, it could be proposed to select large sample sizes sampled from large universes.

- Conducting empirical research highlighting effective procedures in an attempt to reduce sources of error variances, especially the ones associated with students-by-task interaction, which arguably affected the reliability of scores obtained from performance/competence assessments implementing open ended-tasks. Some possible suggestions include investigating parallel forms (tasks), transfer and concept mapping, and altering G study designs.

- Using G theory to study validity and reliability of scores to be gained from other forms of performance/competence assessments, such as self and peer assessments, project works, and observation methods.

References

Afzal, N. (2019). A study on vocabulary-learning problems encountered by BA English majors at the university level of education. *Arab World English Journal, 10*(3), 81-98. DOI: http: //dx.doi.org/10.24093/awej/voi10no3.6.

Agdam, S, J., & Sadeghi, K. (2014). Two formats of word association tasks: A study of depth of word knowledge. *English Language Teaching, 7(10),* 1-12.

Akindahunsi, O. F., & Afolabi, E. R. I. (2021). Using generalizability theory to investigate the reliability of scores assigned to students in English language examination in Nigeria. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 147-162. DOI: 10.21031/epod.820989.

Alduais, A. M. S. (2012). An account of approaches to language testing. *International journal of Academic Research in Progressive Education and Development, 1(4),* 203- 208.

Alkharusi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education, 2* (1), 184-196.

Allem, S. E. M. (2000). *Alternative educational assessment: its theoretical and methodological foundations and field applications.* Cairo: Dar El fikr Elarabi.

Allem, S. E. M. (2000). *Educational and psychological measurement and evaluation.* Cairo: Dar El fikr Elarabi.

Allen, M., & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/COLE.

American Psychological Association, American Educational Research Association and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Psychological Association, Educational Research Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Psychological Association. (2020). *Publication manual of the American psychological Association: The official guide to APA styles.* (7th edition). Washington, DC: APA.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie, *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, DE: International Reading Association.

Andrade, Ch. (2018). Internal, external, and ecological validity in research design, conduct and evaluation. *Indian Journal of psychological medicine*, *40*(5), 498-499.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 19-32). Lawrence Erlbaum Associates, Inc.

Anthony, L. (2020). Resources for researching vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 561-591). Routledge.

Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). London:

Arter, J. (2002). Rubrics, scoring guides, and performance criteria. In C. Boston (Eds.), *Understanding scoring rubrics: A guide for teachers* (pp. 21-31). ERIC: Clearinghouse on Assessment and Evaluation, University of Maryland.

Assessing Writing, 12, 86–107.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van Der Vleuten, C. P. M . (2006). The Wheel of Competency Assessment: Presenting Quality Criteria for Competency Assessment Programs. *Studies in Educational valuation 32*, 153-170.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Evaluating assessment quality in competency-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*, 114-129.

Baartman, L., Gulikers, J., & Asha Dijkstra, A. (2013). Factors influencing assessment quality in higher vocational education. *Assessment & Evaluation in Higher Education, 38*(8) 978-997. Retrieved from: http://dx.doi.org/10.1080/02602938.2013.771133

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191-208.

Babbie, E. R. (2010). *The practice of social research*. Belmont, CA: Wadsworth Cengage.

Bachman, L. F & Palmer, A. S. (1996). *Language testing in practice: Designing and*

developing useful language tests. Oxford: Oxford university press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, 5-18.

Bain, D., & Pini, G. (1996). *Pour évaluer vos evaluations. La Généralizabilité: Mode d'emploi*. Genève: Centre de Recherche Psychopédagogiques. Direction Générale du Cycle d'orientation.

Bardakçı, M. (2016). Breadth and depth of vocabulary knowledge and their effects on L2 vocabulary profiles. *English Language Teaching, 9*(4), 239.-250.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education*, *4*(4), 305–318.

Baxter, G. P., & Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality diverse studenst impact. *Journal for Research in Mathematics Education, 24* (3), 190-216.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation and procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29* (1), 1-17

Beck, I. L., McKeown, M. G., & Omanson, R. C. (1987). The effects and uses of diverse vocabulary instructional techniques. In M.G. McKeown & M.E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 147–163). Hillsdale, NJ: Erlbaum.

Bertrand, R., & Blais, J. G. (2004). *Modèles de mesure: L'apport de la théorie des réponses aux items*. Canada: Presses de l'Université du Québec.

Bolus, R., Hinofotis, F., & Bailey, K. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*(2), 245–258.

Bon, T. G., & Fox, C. M. (2015). *Applying the rash model: Fundamental measurement in the human sciences.* New York, NY: Routledge/Taylor and Francis Group.

Bottema-Beutel, K., Lloyd, B., Carter, E. W., & Asmus, J. M. (2014). Generalizability

and decision studies to inform observational and experimental research in classroom settings. *AMERICAN JOURNAL ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES*, *119* (6), 589–605.

Brennan, R. L. (1992 c). An NCME instructional model on Generalizability Theory. Instructional topics in educational measurement, Module 14. Madison, WI: National Council on Measurement in Education.

Brennan, R. L. (1992a). Generalizability theory. *Educational Measurement: Issues and Practice, 11*, 27-34.

Brennan, R. L. (1992b). Elements of generalizability theory (Rev. ed.). Iowa City, IA: American College Testing Program.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement*: *Issue and Practice, 16* (4), 14-20.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Brennan, R. L. (2010). Generalizability theory. *Educational Measurement*, 61-68, in: http://www.education.uiowa.edu.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14* (3), 259-277.

Brennan, R.L. (2000). Performance Assessments from the Perspective of Generalizability

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014) Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of Psychology, 52*, 13-35.

Brookhart, S. M. (1999). The art and science of classroom assessment: The missing part of pedagogy. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. Second Language Studies, *22*(2), pp. 91-139.

Brown, T, Bonsaksen, T, & Hui, F. K. F. (2019). An examination of the structural validity of the physical self-description questionnaire-short from (PSDQ-S) using the rash measurement model. *Cogent Education, 6* (1), 1-28.

Brualdi, A. (2002). Implementing performance assessment in the classroom. In C. Boston (Ed.), *Understanding scoring rubrics: A guide for teachers* (pp.5-14). ERIC Clearinghouse on Assessment and Evaluation: University of Maryland.

Burt, C (1955). The evidence for the concept of intelligence. *British Journal of Educational Psychology, 25* (3), 158-177. https://doi.org/10.1111/j.2044-8279.1955.tb03305.x.

Burt, C. 1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *the marks of examiners* (pp. 245-314). London: Macmillan.

Burton, R. C. (2008). Oral retelling as a measure of reading comprehension: The generalizability of ratings of elementary school students reading expository texts [Master thesis, Brigham Young University]. From: https://scholarsarchive.byu.edu/etd/1678 google.com.

C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316 –223). Clevedon, UK: Multilingual Matters.

Cardinet, J., Jhonson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. Routledge: New York.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Application to educational measurement. *Journal of Educational Measurement, 13*(2), 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18* (4), 183-204.

Castle, J. (2018). *Performance-based assessment* [Prezi slideshow presentation]. https://www.slideshare.net.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. Language Testing, 12(1), 16-33.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.), Researching pedagogic tasks: Second language learning, teaching and testing (pp. 210-228). Harlow, UK: Pearson Education.

Chomsky, N. (1965). *Aspects of the theory of syntax.* M.I.T. Press

Clauser, B. E., Subhiyah, R. G., Nungester, R. J., Ripkey, D. R., Clyman, S. G., & McKinley., D. (1994). Scoring a performance-based assessment by modelling the judgements of experts. *Journal of Educational Measurement, 23* (4), 397-415.

Clemans, W. (1971). Test administration. In R. Thorndike (Ed.), *Educational measurement* (pp. 188-201). Washington, DC: American Council on Education.

Cobb, T. (2006). *VocabProfiler.* Retrieved 15/03/2020, from http://www.lextutor.ca/vp/

Colman, A. M. (2015). *A dictionary of psychology*. USA: Oxford University Press. Competence: From Methods to Programmes. *Medical Education*, 39, 309–317.

Coulacoglou, C & Saklofske ,D.H. (2018). *Psychometrics and psychological assessment: principles and applications*. USA: Elsevier Academic Press.

Council of Europe (CEFR). (2001). *Common European Framework of reference for Languages: Learning, teaching, assessment*. New York: Cambridge University Press.

Crick, J. E., & Brennan, R. L. (1983). Manual for GENOVA: A generalized Analysis of variance system. (American College Testing Technical Bukketin No. 43). Iowa city, IA: American College Testing, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research, 36* (3), 206-217.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16* (3), 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: John Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373-399.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16, 137-163.*

Crystal, D. (1995). Cambridge Encyclopedia of the English Language. Cambridge: Cambridge University Press.

Cumming, A, H. (1996). Validation in Language Testing Modern Languages in Practice. In A. H. Cumming & R. Berwick (Eds.), *Introduction: The Concept of Validation in Language Testing* (pp.1-14). England Philadelphia: Multilingual Matters.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42,* 895-901.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University.

De Gruijter, D. N. M., & Van Der Kamp, L. J. Th. (2005). *Statistical Test Theory for Education and Psychology*. New York: Tylor & Francis Group.

De Gruijter, D. N. M., & Van Der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioral sciences*. New York: Taylor & Francis Group.

De vet, H. C. W., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman-Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology, 85,* 45-49.

Delory, C. (2002). L'évaluation des compétencesdans l'enseignement fundamental: De quoi parle-t-on ? In L. Paquay., Carlier. Gh, & Huynen (Eds), L'évaluation des compétences chez l'apprenant: Pratiques, methods et fondements (pp. 21-35). Louvain : PUL.

Deville, C., & Deville, M. C. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. C. Deville., C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives*

(pp 9-23). John Benjamins Publishing Company: Amsterdam and Philadelphia.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*, 1006-1012.

East, M. (2004). Calculating the lexical frequency profile of written German texts. *Australian Review of Applied Linguistics, 27* (1), 30-43.

Ebel, R. R. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424. http://dx.doi.org/10.1007/BF02288803.

Ebrahimi, A. B. (2017). Measuring productive depth of vocabulary knowledge of the most frequent words. (T. U. Western, Éd.) Ontario, Canada. Retrieved August 8th, 2021, from: https://ir.lib.uwo.ca/etd/4894/

Edmonds, A., Clenton, J., & Elmetaher, H. (2022). Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge. *System*, 1-38.

Ehsanzadeh, J, S. (2012). Depth versus breadth of lexical repertoire: Assessing their roles in EFL students' incidental vocabulary acquisition. *TESL CANADA JOURNAL 29* (2), 24-41.

Elliot, S. N., & Roach, A. T. (2007). Alternative assessments of students with significant disabilities: Alternative approaches, common technical challenges. *Applied Measurement in Education*, 20(3), 301-333.

From: https://doi.org/10.1080/0895701431385

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4* (2), 139–55.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational Measurement (pp. 127-144). New York: Macmillan.

Finlayson, D. S. (1951). The reliability of the marking of essays. *British Journal of Educational Psychology*, *21* (2), 126-134.

Finocchiaro, M., & C. Brumfit. (1983). The Functional-Notional approach: From theory to Practice. New York: Oxford University Press.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society, 22*, 700-725.

Fitzpatrick, R., & Morrison, E. (1971). Performance and product evaluation. In R. Thorndike (Ed.), *Educational measurement* (pp. 237–270). Washington, DC:

American Council of Education.

Fountain, R. L., & Nation, I. S. (2000). A vocabulary-based graded dictation test. *RELC Journal, 32* (2), 29-44.

Fritz, E., & Rueggb, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing, 18,* 173-181. From: https://doi.org/10.1027/1015-5759/a000664.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26* (4), 507–531.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*, 100-117. doi:10.1016/j.asw.2010.05.002.

Genesee, F., & Upshur, J. A. (1996). *Classroom-Based Evaluation in Second Language Education*. Cambridge: Cambridge University Press.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: The Falmer Press.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika 30*, 395–418.

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41* (4), 481-505. https://doi.org/10.1093/applin/amy057.

Gottardo, A., Mirza, A., Koh, P.W., Ferreira, A., & Javier, C. (2017). Unpacking listening comprehension: The role of vocabulary, morphological awareness, and syntactic knowledge in reading comprehension. *Read Writ*, 1-24.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabualry in a Second Language: Selection, acquisition and testing* (pp. 191-208). John Benjamins Publishing Company.

Gren, L. (2018). Standards of validity and the validity of standards in behavioral software engineering research: The perspective of psychological test theory. In *Proceedings of ACM/IEEE International Symposium on Empirical Software*

Engineering and Measurement (ESEM), Oulu, Finland, October, 11-12, 2018 (ESEM'18), 4 pages. DOI: 10.1145/3239235.3267437.

Gulikers, J., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competence-based assessment Quality. *Studies in Educational Evaluation 35*, 110-19. Elsevier.

Gulikers, J.T.M., Bastiaens, T.J., & Kirschner, P.A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research & Design*, 52, 67-87.

Gullikson, H. (1950). *Theory of mental tests*. New York: Wiley Publications.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective: Challenges and potential solutions. Dans C. Bardel, C. Lindquist, & L. Laufer, *L2 vocabulary acquisition, knowledge and use* (pp. 11-28). European Second Language Association: Eurosla Monographs Series.

Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across Multiple Tasks: A Challenge for Authentic Assessment. Research and practice in assessment, 7, 16-28.

Hazenberg, S., & Hulstijn, J. H. (1996). Defining a minimal second language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics, 17*,145-163.

Hedge, J. W., & Teachout, M. S. (2000). Exploring the concept of acceptability as a criterion for evaluating performance measures. *Group & Organization Management, 25*(1), 22-44.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in second language acquisition, 21*(2), 303-317.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). A practical guide to alternative assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

Hill, H., Ogle, K, .Gottlieb, M., Santen, S. A., & Artino A. R. (2022). Educator's blueprint: A how to guide for collecting validity evidence in survey-based research. *AEM Education and training, 6*(6), 1-5 from: https://doi.org/10.1002/aet2.10835.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified

texts for pleasure? *Reading in a Foreign Language, 8*, 689 –696.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160. http://dx.doi.org/10.1007/BF02289270. http://refhub.elsevier.com/S0022-4405(13)00110-6/rf0065. https://doi.org/10.1111/j.1467-1770.1982.tb00970.x.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403-430.

Huang, C. (2009). Magnitude of task sampling variability in performance assessment: A meta-analysis. *Educational and psychological measurement, 69* (6), 887-912.

Huang. J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17* (2012), 123-139.

Hudson & J. D. Brown (Eds.), A focus on language test development: Expanding the language proficiency construct across a variety of tests (Technical Report #21) (pp.163-204). Honolulu, HI: University of Hawai‘i, Second Language Teaching & Curriculum Center.

Iliescu, D., & Greiff, S. (2021). On consequential validity. *European Journal of Psychological Assessment, 37* (3), 163-166.

Im, G. H, Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia*, 1-26. https://doi.org/10.1186/s40468-019-0089-4

Johnson, R. l., Penny, J. A., & Gordon, B. (2009). *Assessing Performance: Designing, scoring, and validating performance tasks*. New York: The Guildford Press.

Jonsson, A., & Svingby,G. (2007) The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review* (2), 130-144.

Jung, I. (2016). A framework for assessing fitness for purpose in open educational resources. *International Journal of Educational Technology, 13* (3), 1-11.

Kamlasi, I., & Nokas, D. N. (2017). Grammatical errors in writing of the second class students of SMA Kristen 1 Soe. *Metathesis, 1* (1), 130-140.

Kan, A .(2007). An alternative method in the new educational program from the point of performance-based assessment: Rubric scoring scales. *Educational Sciences: Theory & Practice*, *7* (1), 144-152.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.

Kane, M. T. (1999). The role of generalizability in validity. Presentation at NCME meeting, ERIC TM 029 888, pp. 1-11. Scholar.google.com.

Karakoç, D., & Köse, G. D. (2017). The impact of vocabulary knowledge on reading, writing and proficiency scores of EFL learners. *Journal of Language and Linguistic Studies, 13* (1), 352-378.

Khodi, A. (2021). The affectability of writing assessment scores: a G-theory analysis of rater, task, and scoring method contribution. *Language Testing in Asia*, 23-27. https://doi.org/10.1186/s40468-021-00134-5.

Koizumi, R. & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4* (5), 900-913.

Koizumi, R. (2013). Vocabulary and speaking. *The Encyclopaedia of Applied Linguistics*, 1-7. Retrieved from: https://doi.org/10.1002/9781405198431.wbeal1431.

Kumazawa, T. (2009). Revision of a criterion-referenced vocabulary test using generalizability theory. *JALT Journal, 31*(1), 81-100.

Kyle, K., & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. , 1-30. *TESOL QUARTERLY, 49*(4), 757-786.

Lado, R. (1961). *Language testing*. London: Longman.

Lai, R.E. (2011). Performance-based Assessment: Some New Thoughts on an Old Idea. *Bulletin*, 20, Pearson Education, Available at: www.pearsonassessments.com.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied Measurement in Education, 2* (3), 195–205.

Lather, P. (1986). Issues of validity in openly ideological research: Between a rock and a soft place. *Interchange 17*, 63-84.

Laufer, B. & I.S.P. Nation. (1999). A vocabulary size test of controlled productive ability. *Language Testing 16*(1), 33-51.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L.

Laufer, B., & I. S. P. Nation. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics 16*(3), 307-322.

Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–76). New York, NY: Routledge.

Laveault. D., & Grégoire. J. (2002). *Introduction aux théories des tests en psychologie et en*

Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G theory. *International journal of testing, 7* (4), 353-385.

Li, M., & Kirby, J. R. (2014). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics,* 1-25.

Lindquist, E. F. (1953). Design and analysis of experiments in phycology and education. Houghton Mifflin.

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. *Eurosla Monographs Series, 2*, 109-126.

Linn, R. (1994a). Evaluating the technical quality of proposed National Examination Systems, *American Journal of Education*, 102 (4), 565-580.

Linn, R. L, Baker, E. L & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20* (8), 15-21.

Linn, R. L. (1994b). Performance assessment: Policy promise and technical measurement standards. *Educational Researcher, 23*, 4-14.

Llabre, M. M. (1978). *An application of generalizability theory to the assessment of*

writing ability. Florida: University of Florida.

Lord, F. M. (1955). Estimating test reliability. *Educational and psychological measurement, 15*, 325-336.

Lord, F. M. (1957). Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement, 17*, 510-521. http://dx.doi.org/10.1177/0013164490504004.

Lord, F. M. (1959). Test of the same length do have the same standard errors of measurement? *Educational and Psychological Measurement, 19*, 233-239. http://dx.doi.org/10.1177/001316445901900208.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley: MenloPark.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rash measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-80.

MacIntyre, N. J., Bennett, L., Bonnyman, A. M., & Stratford, P. W. (2011). Optimizing reliability of digital inclinometer and flexicurve ruler measures of spine curvatures in postmenopausal women with osteoporosis of the spine: an illustration of the use of generalizability Theory. *International Scholarly Research Network*, 1-9.
Macmillan.

Magnusson, D. (1967). *Test theory*. Reading, Mass: Addison-Wesley Publishing Company.

Marcoulides, G. A. (2000). *Handbook of applied multivariate statistics and mathematical modeling*. California: Academic Press.

Marcoulides, G. A., & L. Kyriakides. (2010). Using generalizability theory. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational*. *effectiveness research* (pp 219-245). New York: Routledge.

Martinez, J, F., Goldschmidt, P., Niemi, D., Baker, E. l., & Sylvester, R. M. (2007). Language arts performance assignments: Generalizability studies of local and central ratings. *Educational Assessment, 12* (4), 267-282.

McBee, M. M., & Barns, L. L. B. (1998). The Generalizability of a Performance Assessment Measuring Achievement in Eighth-Grade Mathematics. *Applied*

Measurement in Education, 11 (2), 179-194.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459-488.

McTighe, J., & Ferrara, S. (1998). *Assessing learning in the classroom*. Washington, DC: National Education Association.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition.* Amsterdam: John Benjamins.

Meara, P. M., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System, 28*, 19-30.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 142*(4), 142-154.

Meara. (1996). The dimensions of lexical competence. Dans G. Brown, K. Malmkjaer, & J. Williams, *Performance and competence in second language acquisition* (pp. 35-53). Cambridge University Press.

Menéndez-Varela, J. L., & Gregori-Giralt, E. (2017). The reliability and sources of error of using rubrics-based assessment for student projects. *Assessment & Evaluation in Higher Education*, 1-13. DOI: 10.1080/02602938.2017.1360838.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (pp. 13-103).* New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, *23* (2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist, 50,* 741-749.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256.

Meyer, J. P. (2010). *Reliability*. New York: Oxford University Press.

Miller, M. D. & Linn R. L. (2000). Validation of Performance-Based Assessments.

Applied Psychological Measurement, 24 (4), 367–378.

Miller, M.D. (2010). Classical Test Theory Reliability in International Encyclopedia of Education , 27-30

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Multilingual Matters.

Milton, J., & Hopkins, N. (2005). *AuralLex.* Swansea, Wales: Swansea University.

Ministry of National Education (2006). New Prospects: Programme d'anglais deuxieme langue étrangere. Ministry of National Education. Algiers.

Ministry of National Education. New Prospects. (2017). Ministry of National Education. Algiers.

Miralpeix, I. (2020). L1 and L2 vocabulary size and growth. Dans S. Webb, *The Routledge handbook of vocabulary studies* (pp. 189-206). Routledge.

Moskal, B. M. (2019). Scoring Rubrics: What, When and How? *Practical Assessment, Research, and Evaluation: 7* (3), 1-6. DOI: https://doi.org/10.7275/a5vq-7q66

Muijs, D. (2010). *Doing quantitative research in education with SPSS*. London: Sage Publications.

Mukarto, F. X. (2005). Assessing the depth of second language vocabulary knowledge. *Physics, 8*(3), 151-169.

Nagy, W., & Scott. (2000). Vocabulary processes. In M. L. Kamil., P. B. Mosenthal., P. D. Pearson & R. Barr (Eds.), *Handbook of Reading Research* (pp. 269–284). Mahwah, NJ: Lawrence Erlbaum.

Nathaniel Hawthorne. (n.d.). AZQuotes.com. Retrieved June 20, 2023, from AZQuotes.com Web site: https://www.azquotes.com/quote/127016

Nation, I. S. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, I. S. (2005). Teaching and learning vocabulary. Dans E. Hinkel, *Handbook of research in second language teaching and learning* (pp. 581-596). New Jersey: Lawrence Erlbaum Associates Publishers.

Nation, I. S. (2013). *Learning vocabulary in another language* (éd. 2). Cambridge: Cambridge University Press.

Nation, I. S. (2014). How much input do you need to learn the most frequent 9,000

words? *Reading in a Foreign Language, 26*(2), 1-16.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Boston: Heinle and Heinle.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59-82.

National Association of State Boards of Education. (2020). Performance assessments: promises and pitfalls. www.nasbe.org.

National Capital Language Resource Center (NCLRC). (n.d.). *The Essentials of Language Teaching.* http://nclrc.org/essentials/assessing/alternative.htm.

Norris, J. M. (2001). Identifying rating criteria for task-based EAP assessment. In T.

Norris, J., Brown, J., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments (Technical Report #18). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

Osterlind, S. G. (2002). Constructing test items: multiple-choice, constructed-response, performance, and other formats. Kluwer Academic Publishers: New York.

Oxford Languages and Googles English Dictionary. (2023). Oxford: Oxford University Press. From: languages.oup.com.

Ozturk, G. (2012). The effect of context in achievement vocabulary tests. *Journal of Educational and Instructional Studies in the World, 2* (4), 126-134.

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*(1), 9-29.

Parkes, J. (2000). The interaction of assessment format and examinees' perceptions of control. *Educational Research, 42* (2), 175-182.

Parkes, J. (2001). The role of transfer in the variability of performance assessment scores. *Educational* Assessment, *7* (2), 143-164. http://dx.doi.org/10.1207/S15326977EA0702 3.

Parkes, J. (2010). The interaction of assessment format and examinees' perceptions of control. *Educational Research, 42*(2), 175–182. http://www.tandfonline.com/loi/rere20.

Parkes, J., Zimaro, D. M., Zappe, S. M., & Suen, H. K. (2000). Reducing task-related variance in performance assessment using concept maps. *Educational Research and Evaluation An International Journal on Theory and Practice, 6* (4), 357-378.

Paul, P.V., Stallman, A. C., & O' Rourke, J. P. (1990). Using three test for mats to assess good and poor readers' word knowledge. Technical Report No.509, Center for the Study of Reading, University of Illinois at Urbana-Champain, IL.

Perlman, C. (2002). An introduction to performance assessment scoring rubrics. In C. Boston (Ed.), *Understanding scoring rubrics: A guide for teachers* (pp, 1-4). ERIC Clearinghouse on Assessment and Evaluation: Maryland.

Pilliner (1952). The application of analysis of variance to problems of correlation. *British Journal of Statistical Psychology, 5* (1), 31-38. http://10.1111/j.2044-8317.1952.tb00109.x.

Pini, G. (2010). Relations entre le coefficient de généralisabilité absolu et les indices RhÔ Carré et Omega Carré. Groupe Eudiométrie-Qualité de la mesure en éducation.

Polat, M., & Turhan, N. S. (2021). Applying generalizability theory in language testing: Comparing nested and crossed scoring designs in the assessment of speaking skills. *International Journal of Curriculum and Instruction 13* (3), 3344-3358.

Popham, J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership, 55* (2), 72-75.

Qian, D. D. (1998). Depth of vocabulary knowledge: Assessing its role in adults' reading comprehension in English as a second language. *Unpublished Doctoral Thesis*. University of Toronto. Récupéré sur https://tspace.library.utoronto.ca/handle/1807/12079

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*, 282-308.

Qian, D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21*, 28-52.

Read, J & Nation, P. (1986). Some issues in the testing of vocabulary knowledge. Eric Cleaninghouse. Wachigton DC p. 28 (pages). Paper presented at a language testing symposium. Available at: http://eric.ed.gov/

Read, J. (1989). *Towards a deeper assessment of vocabulary knowledge.* Washington, DC: ERIC Clearing House on Languages and Linguistics.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10,* 355-371.

Read, J. (2000). *Assessing vocabulary.* Cambridge University Press.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary nowledge be defined? In P. Bogaards, & B. Laufer (Eds), *Vocabulary in a second language* (pp. 209-227). Amsterdam: Benjamins.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7*(2), 105-125.

Read, J. (2012). Assessing vocabulary. In C. B. Coomb, P. Davison, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp.257-263). Cambridge: Cambridge University Press.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.

Read, J., & Dang, T. (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research,* 1362-1688 .

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10* (1), 77-89.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26* (1), 108-116.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performnce assessments. *Journal of Educational Measurement, 30* (1), 41-53.

Sanders, A. (1998). *Elements of human performance*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sari, E., & Han, T. (2022). Using generalizability theory to investigate the variability and reliability of EFL composition scores by human raters and e-rate. *Porta Linguarum, 38,* 27-45.

Sasayama, S. (2011). Cognition hypothesis and second language performance: comparison of written and oral task performance. *Second Language Studies, 29* (2),

107-129.

Scallon, G. (2004). L'évaluation des compétences et l'importance du jugement. *Pédagogie Collégiale, 18* (2), 15-20.

Schellekens, L. H., Kremer, W. D. J., Van der Schaaf, M. F., Van der Vleuten, C. P. M., & Bok, H. G. J. (2023). Between theory and practice: Educators' perceptions on assessment quality criteria and its impact on student learning. *Frontiers in Education, 08*, 1-9.

Schmidgall, J. E. (2017). *The consistency of TOEIC speaking scores across ratings and tasks* (Research Report No. RR-17-46). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12178

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*, 281-317.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *SSLA, 20*, 17-36.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22* (1), 1–30.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effects of rater, task and language. In E. Van Steendam, M. Tillima, G. Rijlaarsdam, & H. Van Den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices studies in writing(*pp.1–22). Emerald Group Publishing Limited: Leiden Boston , the Netherlands.

sciences de l'éducation. Bruxelles: De Boeck.

Sharakhimov, S., & Nurmukhamedov, U. (2021). Assessing learners' productive vocabulary knowledge: Formats and considerations. *English Teaching Forum, 59*(4), 16-25.

Shavelson, R. G., Baxter G. P., & Gao, X. (1993). Sampling variability of performance Assessments. Journal of Educational Measurement, 30(3), 215-232.

Shavelson, R. G., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political, rhetoric and measurement reality. *Educational Reasearcher, 21*(4), 22-27.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* USA: Sage Publications, Inc.

Shavelson, R. J., & Webb. N. M. (2009). Generalizability theory and its contribution to the discussion of the generalizability of research findings. In K Ercikan, & W. M. Roth (Eds.), *Generalizing from educational research: Beyond Qualitative and Quantitative Polarization* (pp.13-32). Routledge: London and New York.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher. 29* (7), 4-14.

Shumate, S. R., Surle, J., Johnson, R. L & Penny, J. A. (2007). The effect of the number of scale points and non-normality on the generalizability coefficient: A Monte Carlo Study, *Applied Measurement in Education, 20* (4), 376-357. http://dx.doi.org/10.1080/08957340701429645.

Skehan, P. (1998a). *A cognitive approach to language learning*. Oxford, UK: Oxford University.

Skehan, P. (1998b). Task-based instruction. Annual Review of Applied Linguistics, 18, 268-286.

Smit, R., & Birri, T. (2014). Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation, 3*, 5-13.

Spearman, C. (1907). Demonstration of formulae or true measurement of correlation. The Aerican Journal of Psychology, 18 (2), 161-169.

Stahl, K. (2018).Teaching academic vocabulary across all content areas vocabulary assessment, 1-33. From: https://nysrti.org.

Stahl, K. A., & Bravo, M. A. (2010). Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher, 63* (7), 566-578.

Swart, M. N., Muijselaar, M. M. L., Steenbeek-Planting, E. G., Droop, M., de Jong, P.F., & Verhoeven, L. (2016). Differential lexical predictors of reading comprehension in fourth graders. *Read Writ, 30* (3), 489-507.

Swiss Society for Research in Education Working Group. (2010). *EduG User Guide*.

IRDP. Neuchatel. Switzerland. In: http://www.irdp.ch/edumetrie/logiciels.html.

Teng, F. (2014). Assessing the Depth and Breadth of Vocabulary Knowledge with Listening Comprehension. *PASAA. 48*, pp. 29-56.

Teng, F. (2016). An in-depth investigation into the relationship between vocabulary knowledge and academic listening comprehension. *The Electronic Journal for English as a Second Language, 20*(2), 1-17.

Thékes, I. (2018). *Studies in the assessment of foreign language vocabulary.* Gál Ferenc College: Gerhardus Publishing.

Theory. *Applied Psychological Measurement, 24 (4), 339–353.*

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273- 286.

Tobar, D. A., Stegner.A. J., & Kane, M. T. (1999).The Use of Generalizability Theory in Examining the Dependability of Scores on the Profile of Mood States. *Measurement in Physical Education and Exercise Science, 3(3), 141-156.* http://dx.doi.org/10.1207/s15327841mpee0303_2.

Van der Vleuten, C. P. M., & L. W. T. Schuwirth. (2005). Assessing Professional

Wang, X. (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL*, *9*, 65-88.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000).The Dependability and Interchangeability of Assessment Methods in Science. *Applied Measurement in Education, 13*(3), 277-301. http://dx.doi.org/10.1207/S15324818AME1303_4

Webb, N. M., Shavelson, R. J., & Haertel, H. E. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics*, *26*, 4-44.

Webb, S. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern language Review, 65*, 441-470.

Webb, S. (2013). *Depth of vocabulary knowledge*. Retrieved from: 4th, 12, 2021, from Wiley Online Library:https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal1325

Webb, S. (2020). *The Routledge handbook of vocabulary studies.* Routledge.

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal,*

44(3), 263–277.

Wesche, B. (1985). Cultural bridges test. Unpublished paper. Monterey Institute of International Studies, Monterey, California.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*(1), 13-40.

Whimbey, A., Vaughan, G. M., & Tatsuoka, M. M. (1967). Fixed effects vs. random effects: Estimating variance components from mean squares. *Perceptual and Motor Skills, 25*, 668.

Wiggins, G. P. (1998). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass Publishers.

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), Implementing performance assessment: Promises and challenges (pp. 61-89). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilkins, D. (1972). *Linguistics in language teaching*. London: Edward Arnold.

Wright, B. D., & Douglass, G. A. (1986). The rating scale model for objective measurement. *MESA Research memorandum, 35*, 1-35.

Yanagisawa, A., & Webb. S. (2020). Measuring depth of vocabulary knowledge. In S, Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371-386). Routledge.

Yashim, A. U., Mhab, L. C., &Waziri, J. A. (2021). Measurement errors in educational assessment. Journal of Educational Theory and Practice (JETP) Online Journal, 1(1), 1-9.

Yelboğa, A., & Tavşancil, E. (2010). The examination of reliability according to Classical test and generalizability on a job performance scale. *Educational Sciences: Theory and Practice, 10* (3), 1847-1854.

Yin, Y., & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *APPLIED MEASUREMENT IN EDUCATION, 21*, 273-291.

Zhang, D., & Yang, X. (2016). Chinese L2 learners' depth of vocabulary knowledge

and its role in reading comprehension. *Foreign Language Annals*, xxx (xxx), 1-17

Zhong, H., F. (2018). The relationship between receptive and productive vocabulary knowledge: A perspective from vocabulary use in sentence writing. *The Language Learning Journal, 46* (4), 357-370.

Notes

1. Note that the terms vocabulary and lexis have been used interchangeably throughout the current research. Crystal (1995) equates vocabulary with lexis or lexicon; this view is to be adopted throughout our research.

2. Depth of vocabulary knowledge and deep word knowledge are to be used interchangeably throughout the work.

3. The term modelling is borrowed from Deller's et al. (2007) volume.

4. The validation process needs various types of evidence, analyses and interpretation in order to establish the link between the test assumptions and the evidence in support of these assumptions (Cumming, 1996).

5. Note that in univariate G theory there exist only one universe of generalization (universe score) per each object of measurement.

6. A statistical procedure set for data analysis obtained from more than one type of observation and measurement involving more than one variable at a time.

7. Lindquist discussed experimental design in details and stressed such issues as precision of measures, randomization, testing hypothesis, complex designs, … etc.

5. It is necessary to go back to the principles of CTT to explain the usefulness of G theory contrasted to CTT, espectially in terms of calculating G coeffcients and providing estimates of the reliability indices associated with underfferntiated error variance.

6. "n" denotes the variance component corresponding to the interaction between the source of variance and the fixed facet, where n, refers to the number of conditions of the fixed facets.

7. Note that the sixth question is related to neither studies because its answer relies on Messick's argument based framework, which depends on collecting validity evidence that can be drawn from an interpretation of both statistical and non- statistical findings, and this will be considered further in the coming chapter.

APPENDICES

Appendix A: Characteristics of Performance Assessments

1. Characteristics of the tasks and characteristics of required processes from the students

1.1. Herman, Aschbacher, & Winters (1992, p. 6) provide a list of performnce assessment characteristics:

1. Tap higher-level thinking and problem-solving skills

2. Use tasks that represent meaningful instructional activities

3. Invoke real-world applications

4. People, not machines, do the scoring, using human judgment

5. Require new instructional and assessment roles for teachers

The five characteristics tap the relationship between the task input and the expected response

2. Characteristics in favour of task specifications

2.1.For Skehan (1998b) characteristics of tasks or content in task-based performance testing involve:

1. Task should be meaningful

2. Task should be communicative containing problem solving situation

3. Task should be authentic; must have relationship to real-word activities

4. Task completion has some significance

5. Assessment of the task is in terms of successful outcome

Skehan's characteristics clearly consider the quality of task content

2.2. Khattri, Reeve, and Kane (1998) propose five criteria to be implemented in designing performance tests:

1. Performance tasks are time demanding

2. Tap cognitive skills by applying problem-solving tasks

3. Tap metacognitive skills (the extent to which learners are aware of their problem solving skills and thinking processes)

4. Social competencies demands (interpersonal skills necessary for successful task completion)

5. Student control dmands (students' ability to define and perform the task successfully)

It sounds clear that these criteria stress the dimensions of task specification that are important in test design and development.

2.3. Norris, Brown, Hudson, and Yoshioka (1998, p. 9-10) also stress other criteria for task design and development besides those mentioned below in the scoring section:

1. Be as authentic as possible with the goal of measuring real-world activities;

2. Sometimes have collaborative elements that stimulate communicative interactions;

d. Be contextualized and complex;

3. Integrate skills with content;

4. Be appropriate in terms of number, timing, and frequency of assessment; and

5. Be generally non-intrusive, i.e., be aligned with the daily actions in the language classroom.

These characteristics dig up task specifications

2.4. Chalhoub-Deville (2001, pp. 214-217) list the following task development characteristics:

1. Task must meet learner centred traits: it should promote individual expression and stimulate learners' prior knowledge and experience

2. Context dependence: task content should be contextualized in on-going discourse and in meaningful situations

3. Authenticity: Task content should bridge the gap between instructional activities and real word activities targeting language use

She emphasizes task content development, and how a task content should look like.

3. Characteristics of the scoring process

3.1. Norris, Brown, Hudson, and Yoshioka (1998, cited in Brown, 2004, p. 21) assert that performance testing depend characteristically on tasks that necessarily need to be weighed by judges relying on certain rating scales:

1. The task should be based on needs analysis (including student input) in terms of rating criteria, content, and contexts

2. The rating scale should be based on appropriate:

a. Categories of language learning and development

b. Appropriate breadth of information regarding learner performance abilities

c. Standards that are both authentic and clear to students

3. To enhance the reliability and validity of decisions as well as accountability, performance assessments should be combined with other methods for gathering information (e.g., self-assessments, portfolios, conferences, classroom behaviors, etc.)

Clearlyy, these criteria are to be invested in rating test takers' performances.

Appendix B: Language Outcomes for Word Building Processes

1- Exploring the past: Ancient Civilizations

Adjectives +preposition (e.g., good at dependent on)

Verb+ preposition (e.g., believe in)

Negative prefixes: dis- and –de

Suffixes: -tion, -ment, -ed, -able, -ic and –ty

Using the past -ed

2- Ill-gotten Gains Never Prosper: Ethics in Business

*Suffix –ty. Eg., honest- hones**ty**, responsible, responsibili**ty***

Prefixes dis- il- e.g. legal- illegal, honest- dishonest, approve- disapprove

3- Schools Different and Alike: Educational in the World

Forming adjectives with –al and –ive. E.g. educational, reflexive, innovative, responsive, constructive, effective…

Forming nouns: verb+ ing

Forming nouns with –tion

Collocation: schoolmate

4- We Are A Family: Feelings and Emotions

Forming adjectives from nouns with ous, - ful, -ic E.g. courage – courageous, faith - faithful
Forming new words with self- - E.g. self-centre
Forming nouns with – ness and –ty E.g. kind – kindness, loyal-loyalty
Forming verbs with –en, e.g. tight – tighten.

Appendix C: A Checklist of Content Validity of Assessment Tasks

Dear Teachers,

I would be grateful if you could fill in the checklist in your hands. It contains eight communicative situational problem-solving tasks that aim at assessing students' vocabulary knowledge (word meaning, word formation, and word use) of a number of particular words. These tasks are assigned to the baccalaureate holders, who are newly enrolled in the Department of English at ENSB (Ecole Normale Superiure de Bouzareah) stretching from literary streams.

The objective of the study does not only focus on depicting the learner's lexical competence as part of their language proficiency, it also aims at ensuring reliability and validity of test scores that will be obtained from the test tasks. The eight tasks are believed to be appropriate in reflecting a number of criteria (parameters) for assessing the lexical competence of the participants. These tasks are used as a means to assess the extent to which the conditions that are necessary for vocabulary performance assessment tasks have been met.

Please, put a cross (**x**) in every column next to "**Yes**" or "**No**" answers related to every task. If your answer is "Yes" this means that the condition is suitable for the situation (task). If you answered "No" please put a comment or suggestion below the checklist for further modification.

Note: please, keep the tasks away from the students' eyes because they can be the sample of this study. If so, these tasks will not serve the purpose of the study and will negatively affect test results.

Thank you in advance for your cooperation!

The Researcher: *WASSILA TEBAA*

A Checklist of Construct Validity

Parameter	Yes	No
1. The task matches the objective (measuring student's vocabulary knowledge).		
2. The task integrates knowledge, skills and attitudes rather than a simple recall.		
3. The task is a real life-like activity (reflects real life language use).		
4. The task stimulates students' thinking skills.		
5. The task is intended for assessing students' performance.		
6. The task allows students to respond differently.		
7. The task contains the necessary information needed to arrive at the correct answer.		
8. The task is structured so that students can have control over response format.		
9. The task structure does not favour some students at the expense of others.		
10. The task is of appropriate level of difficulty for students.		
11. The task (topic and content) addresses positive values and goals relevant to the students, society and education systems.		
12. Task content is not sensitive (not hurting or displeasing).		
13. The task includes cultural content appropriate to the students.		
14. Task content allows students to know how the response will be evaluated.		
15. Task instruction is clear.		
16. The content of the task is understandable.		
17. The task is familiar (practiced previously) but its content is new to all students.		
18. The task is interesting to the students.		
19. Task situation contains the basic components (context, prompt, and instruction)		
20. The language of the task is free from lexical redundancy.		
21. Time allocated for the task is sufficient.		

Other suggestions, or modifications, please specify!

……

……

……

……

……

……**Than thank you very much for your cooperation!**

Appendix D: Language Test for Piloting

Dear students,

This test is a part of a research work. It tends to assess your vocabulary knowledge of some words you have been exposed to in your terminal classes, specifically presented in your textbook "New Prospects" (3éme AS). This test focuses on your ability to use words in appropriate contexts with suitable meanings and word forms.

You are kindly invited to freely answer the following tasks. Test results will remain confidential and will serve the research purposes.

Note: at the end of these tasks, fill in a follow up questionnaire. When answering, please circle the selected answer (Yes/No), or make full sentences when commenting.

Thank you in advance for your corporation.

Task One:

You have been told a story by your grandfather about ancient Egyptian/Sumerian/Ottoman civilization that he has read about Egyptian/Sumerian/Ottoman culture. You wanted to share the story with your classmates! Recite it talking of one civilization, accounting imaginary or real events, **using** the **nouns** of the following **words**: flourish, invent, rise along, fall into ruins.

……
……
……
……
……
……

Task Two:

You listened to a radio documentary about Algerian public revolutions in different regions just after 1830 fighting against French colonization led by famous Algerian leaders. You have been questioned by your teacher of history to write a short paragraph giving a short account about the characteristics underlying ancient Algerian leaders.

Select **four** (04) words that best describe those heroes or their civilization using the adjectives of these words: courage, decline, peace, warlike, knowledge.

……
……
……
……
……
……..

Task Three:

You have been met to a new pupil (or a face booker) unfamiliar with your school who is coming from other English countries. Describe your classroom or school to him/her in terms of location or curriculum studies. Constitute the sentences **using** only four (04)

verbs of these words: situation, graduation, education, qualification, assessment, and training.

……
……
……
……
……
……..

Task Four:

Imagine that you are an experienced teacher and you are asked to plan an ideal school. You will present your plan in a conference attended by UNICEF members in an international meeting. Decide which sort of school it would be. **Use** the **adjectives** of the Following words to describe you plan: discipline, attendance, examination, population.

……
……
……
……
……
……

Task Five:

Suppose that you were a businessman working on mobile phones company. You suffer from many counterfeiters infringing (imitating) your copyright. Write a short memo in which you remind them that imitating property is theft and can cause great deal of financial loss.

Choose **four** (04) words you think will support your ideas. Use the *"ing"* form of these words: counterfeit, swap, theft, financial (loss), consumers. deceive, advertise.

……
……

……
……
……
……..

Task Six:

Imagine you are Human Rights activist against children labor. You have been asked to make an appeal to address people about the need to eradicate this malpractice.

Select **four** (04) words from the list to write a short public statement: unethically, ethical behavior, criminal organization, boycott, violence, exploitation.

……
……
……
……
……
……..

Task Seven:

Suppose that you were a school psychologist, and you are asked to lecture a conference in which you give a piece of advice to a student suffering from stress; being so stressed because of the Baccalaureate examinations.

Choose **four** (04) words stated underneath to achieve success and relief; transferring the words from **adjectives** to **nouns**: self-satisfaction, humour, optimism, worry, fun, stress.

……
……
……
……
……
……..

Task Eight:

You have been told a comic or tragic or love story by a friend and you also want to tell him/her a story or a scene you watched in a film or read in a book about a recent comic or tragic, or love story.

Describe one scene that deeply affected you and express your feeling using four (04) of these words: sadness, fear, grief, love, anger, dislike, funny, irritated, relaxed, happy, crying, proud, satisfied.

……
……
……
……
……
……..

Good Luck!

Appendix E: Survey Questionnaire for Students

Dear students

You are kindly invited to fill in this questionnaire. It aims to collect information about the test content coverage and alignment with test purpose (validity). It tends to examine some other quality criteria of test development such as cognitive complexity, instruction clarity, authenticity, meaningfulness, instructiveness, time allotment and spacing sufficiency for task completion, and topic determination.

Please put a circle on the appropriate answer.

Thank you in advance for your cooperation!

I. Personal information

Name: what is your name?

Age:

Stream (in Secondary School):

Sex:

Region:

II. Questions: Please circle the appropriate answer for each task (*Question items 1 through 10 are used to collect information about each of the eight tasks, and Question items 11 to14 are used to obtain data about the whole test.)*

1- Is the task difficult? *Yes* *No*

2- Are the words of the task complex? *Yes* *No*

If yes, please underline the difficult ones.

3- Is this problem situation new for you? *Yes* *No*

4- Does the task stimulate your thinking? (challenging) *Yes* *No*

5- Is the task realistic? *Yes* *No*

6- Is the task meaningful? *Yes* *No*

7- Is the task interesting? *Yes* *No*

8- Does the task encourage you practice the language you acquired previously?

Yes *No*

9- Is the instruction clear? *Yes* *No*

10-Does the content of the task reflect what you have learned in 3rd year secondary school (knowledge, skills, and attitudes)? *Yes* *No*

Questions on the whole test

11- Is time allocated sufficient for task completion? Yes No

12-Is space left sufficient? Yes No

13- Can you have control over response format? Yes No

14- Can you determine the topic of the task? Yes No

Comments: Other suggestions? Please specify.

……
……
……
……
……
……
……
……

Thank you for your cooperation!

Appendix F: Language Test 2 (for Final Administering)

Time Allocated: 2hours

Dear students,

This test is a part of a research work. It tends to assess your vocabulary knowledge of some words you have been exposed to in your terminal classes, specifically presented in your textbook *"New Prospects"* (3éme AS). This test focuses on your ability to use words in appropriate context with suitable meanings and word forms.

You are kindly invited to freely answer the following tasks. Test results will remain confidential and will serve research purposes.

Thank you in advance for your corporation.

Part I: Personal Information

Name: ……

Age: ……

Gender (sex): ……

Region: ……

Stream in Secondary School: ……

BAC general average: ……

BAC average in English: …….....

Part II: The Test:

Please answer the following tasks concisely and precisely. If you leave one task with no answer this means that your exam sheet will be eliminated from the present study.

Task 1: (5 pts.)

You have read in a book of history about ancient Egyptian/Sumerian/Ottoman civilization or any other civilizations in the world. You wanted to share the story (the information you read about) with your classmates in a project work. Talk of one civilization and narrate its imaginary or real events.

*Select **four** (04) words from the list below and **use** their **verbs** (the past tense) **to explore** some aspects of this civilization: **flourish, invention, rise (along), fall into ruins, contribution,** and **collapse.***

……
……
……
……
……
……
……
……
……

Task 2: (5 pts.)

You listened to a radio documentary about the well-known public revolutions occurred in different regions of Algeria just after 1830. In their attempt to fight against French colonization, these public revolutions were organized under the leadership of many famous Algerian leaders that share in common many characteristics.

You have been asked by your teacher of history to give a short account about the characteristics underlying those ancient Algerian leaders.

*Select **four** (04) words from the following list and **use** their **adjectives** to better **describe** the Algerian heroes or their civilization: **knowledge, peace, nomad, war, courage, ignorance.***

……
……

……
……
……
……
……
……
……

Task 3: (5 pts.)

You have been introduced to a new pupil (or a face booker) unfamiliar with your school who is coming from other English countries. This new friend presented his/her school to you and you were also eager to introduce your school (secondary school or ENS) to hem/her.

Write **"a welcome to our school**" in which you **describe** your school or classroom to your new friend in terms of curriculum studies, or in terms of school system and education procedures.

*Select only **four** (04) words from the following list and **use** their **adjectives** to support your description: **compulsiveness, attendance, education, qualification, assessment,** and **training.***

……
……
……
……
……
……
……
……
……

Task 4: (5 pts.)

You are now a newly recruited student in the teacher training school (ENS); being so means that you gained a lot of experience related to teaching and learning processes along your primary, middle and secondary education.

Imagine that you are invited to an international meeting attended by UNICEF members and you are asked to write a speech in which you suggest a plan for "an ideal school". Based on your experience, decide what sort of school it should be! And how should it function adequately in society to rise good citizens.

*Select **four** (04) words from the list and **use** their **adjectives** to describe your plan:* ***discipline, educate, innovate, construct, affect, and train.***

……
……
……
……
……
……
……
……
……

Task 5: (5 pts.)

Suppose that you have been sold an *"oppo"* smart camera phone that costs between 30 000 to 70 000 DA. Lately, however, you discovered that it is an imitated property. You felt upset and you decided to make an appeal, and publicized in your facebook page, to address businessmen or counterfeiters working on mobile phone companies infringing (imitating) copyrights.

*Choose **four** (04) words from the list and **use** their **noun** form or the **"ing"** form to remind counterfeiters that imitating property is theft and can cause great deal of financial loss*: **steel, consume, counterfeit, deceive, advertise,** and **defraud.**

……
……
……
……
……
……
……
……
……

Task 6: (5 pts.)

You watched in a TV program reportage told about child labor and you were deeply irritated by the fact that managers of companies exploit children with very low wages. You wanted to act like a human rights activist! Write a short public statement in which you convince people, especially employers, about the need to eradicate this malpractice.

*Select **four** (04) words from the list and **use** their **adjectives** to address your audience:* ***unethicality, ethics, crime (organization), boycott, violence,*** *and* ***exploitation.***

……
……
……
……
……
……
……
……
……

Task 7: (5 pts.)

Since you have recently passed your Baccalaureate examinations; you inevitably suffered from stress and anxiety. But you surely managed to overcome them and you

succeeded to get your BAC with high average and now by no means you are a first year university student.

Suppose that one of your relatives, who is this year BAC candidate, asked you to give him/her advice to overcome negative effects of stress to achieve success and relief.

*Choose **four** (04) words stated underneath and **use** their **nouns** to help your relative successfully learn and feel happy: **self-satisfied, humorous, optimistic, worried, funny,** and **stressful.***

……
……
……
……
……
……
……
……
……

Task 8: (5 pts.)

You have been told an exciting comic or tragic story by a friend and you also wanted to tell him/her a story or a scene you watched in a film or read in a book about a recent comic or tragic story.

Describe** one scene that deeply affected you and express your feeling **using** four **(04) adjectives** of these words: **sadness, grief, crying, anger, dislike, irritation, relief, relaxation, happiness, pride, satisfaction,** and **laughter.

……
……
……
……
……
……

……

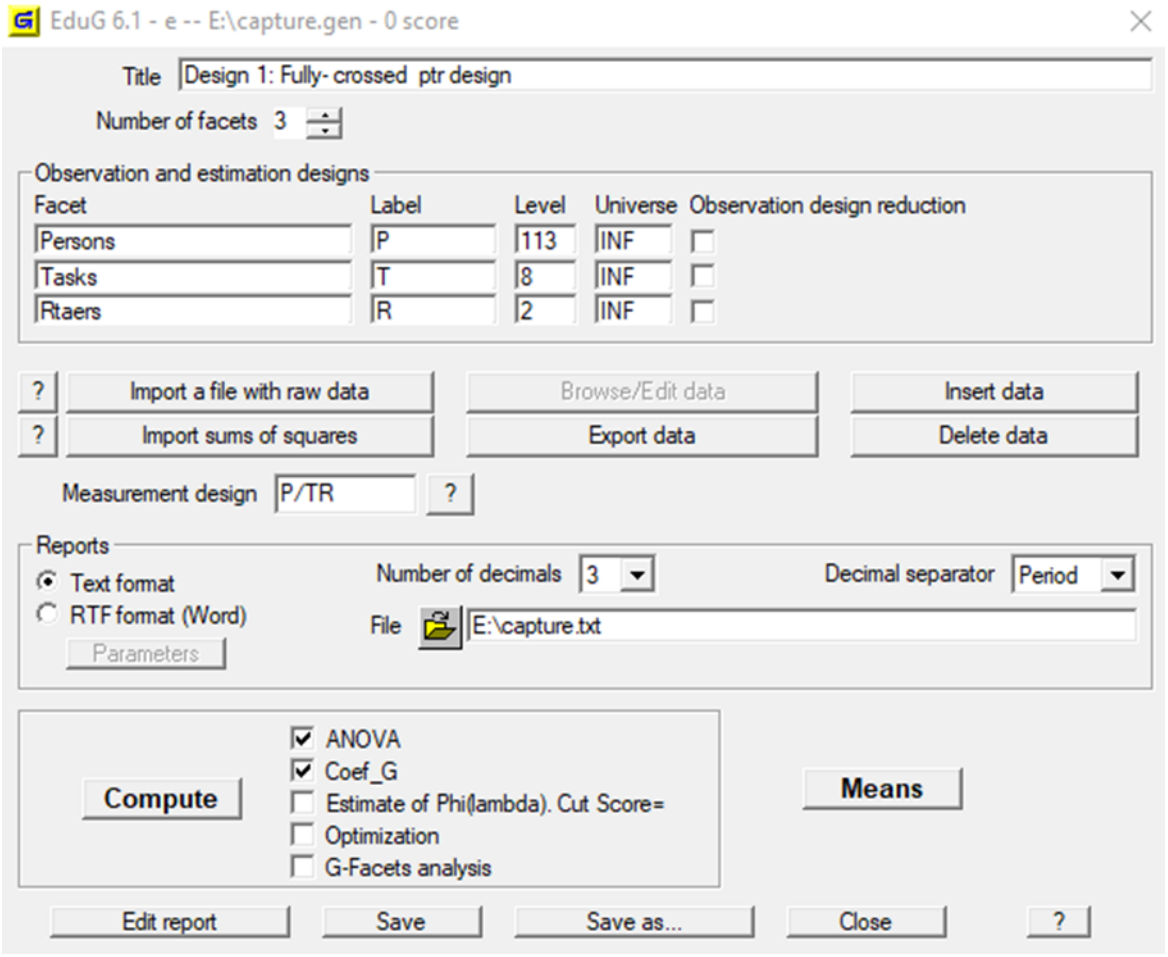
……

……

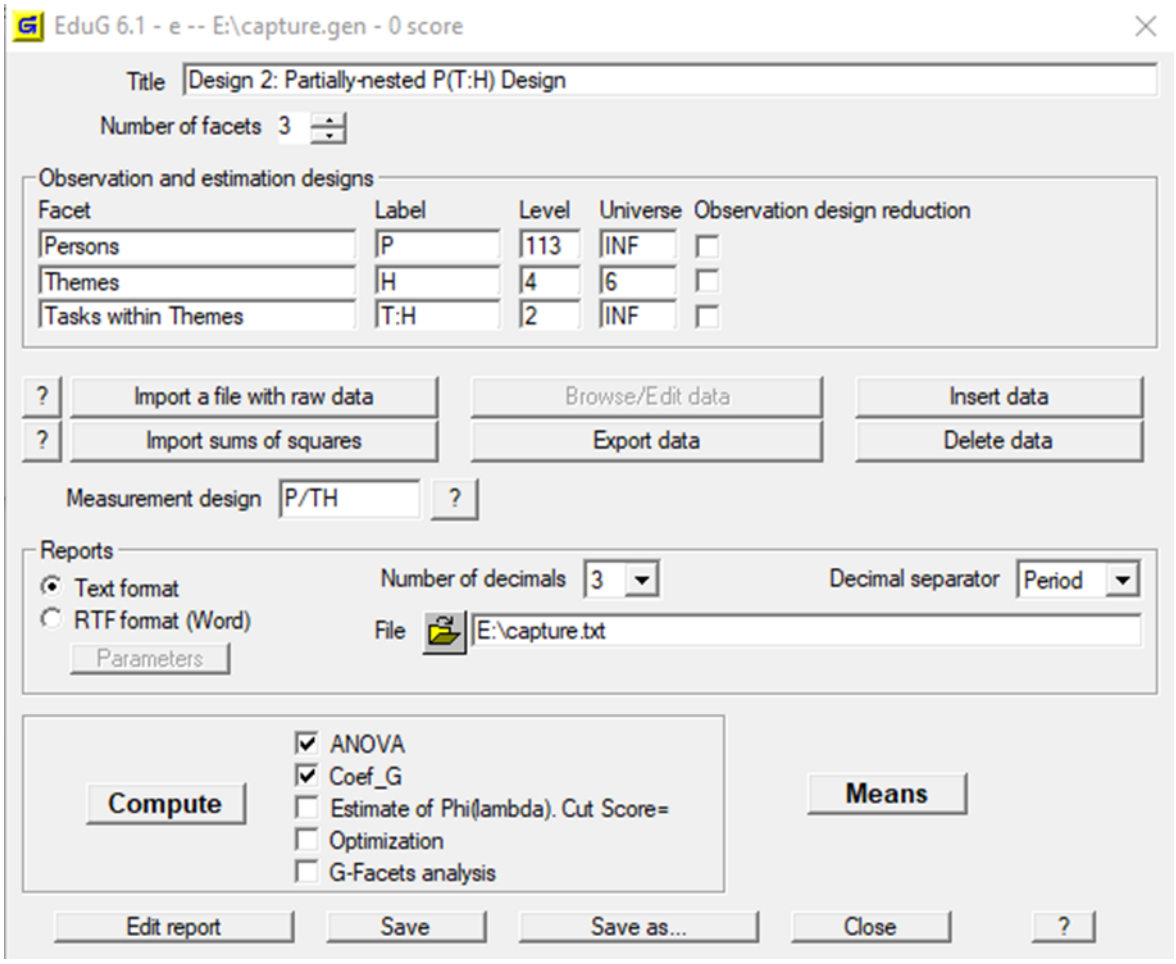
Thank you very much for your cooperation!

Appendix G: EduG Work Screens (Generalizability Studies and Decision Studies)

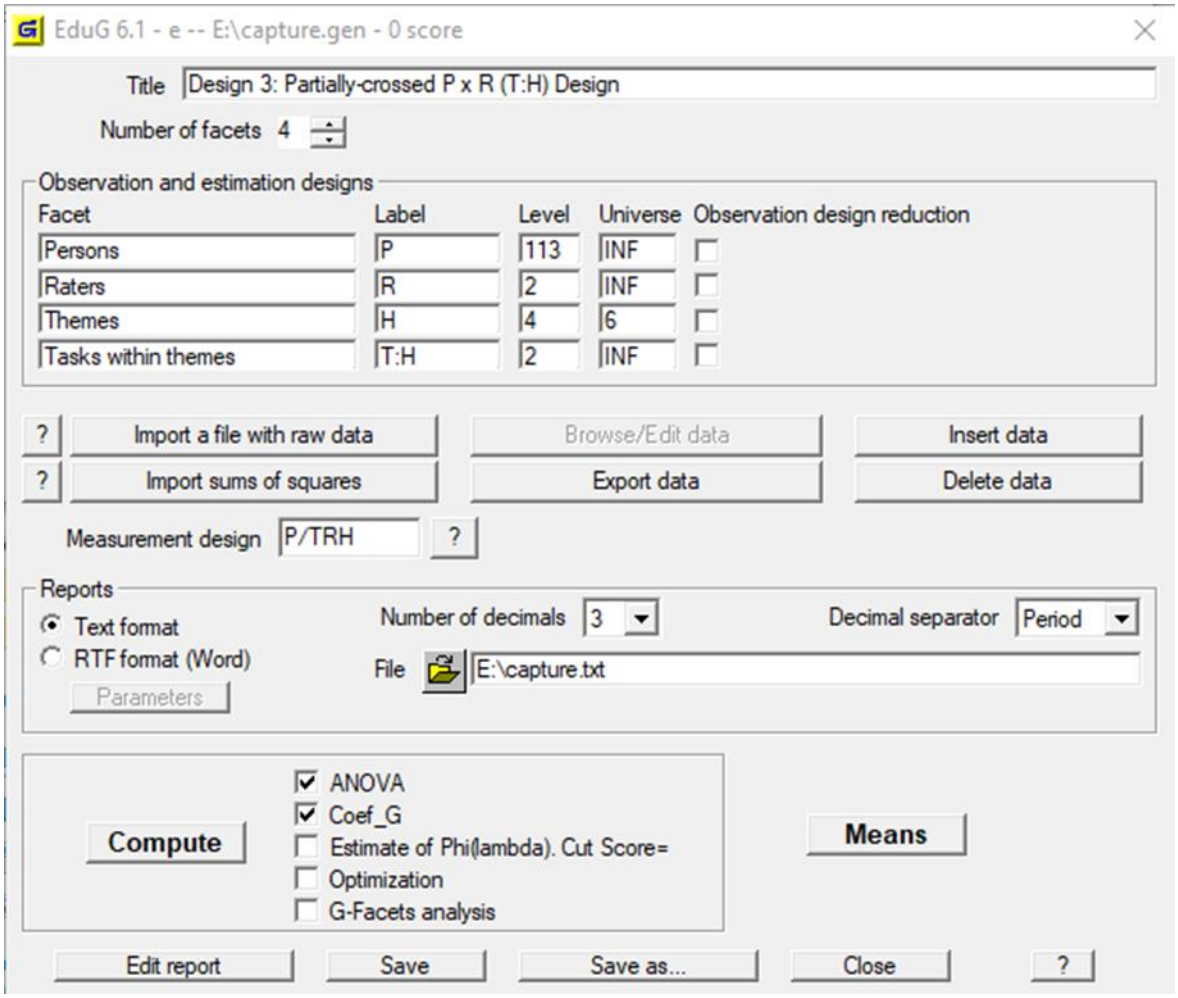
Design 1: Fully-crossed ptr Design



Design 2: Partially-nested P(T:H) Design

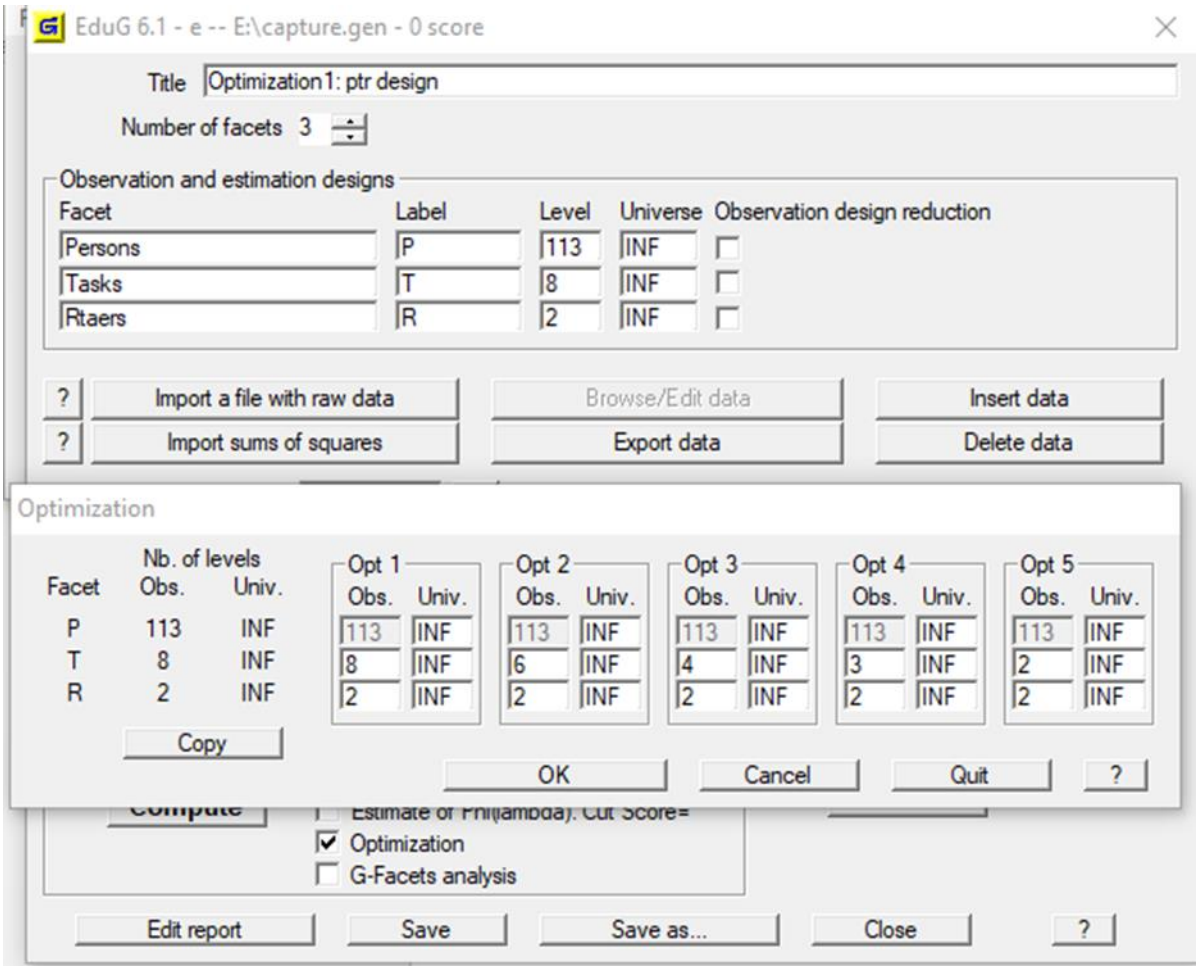


Design 3: Partially-crossed P x R (T:H) Design

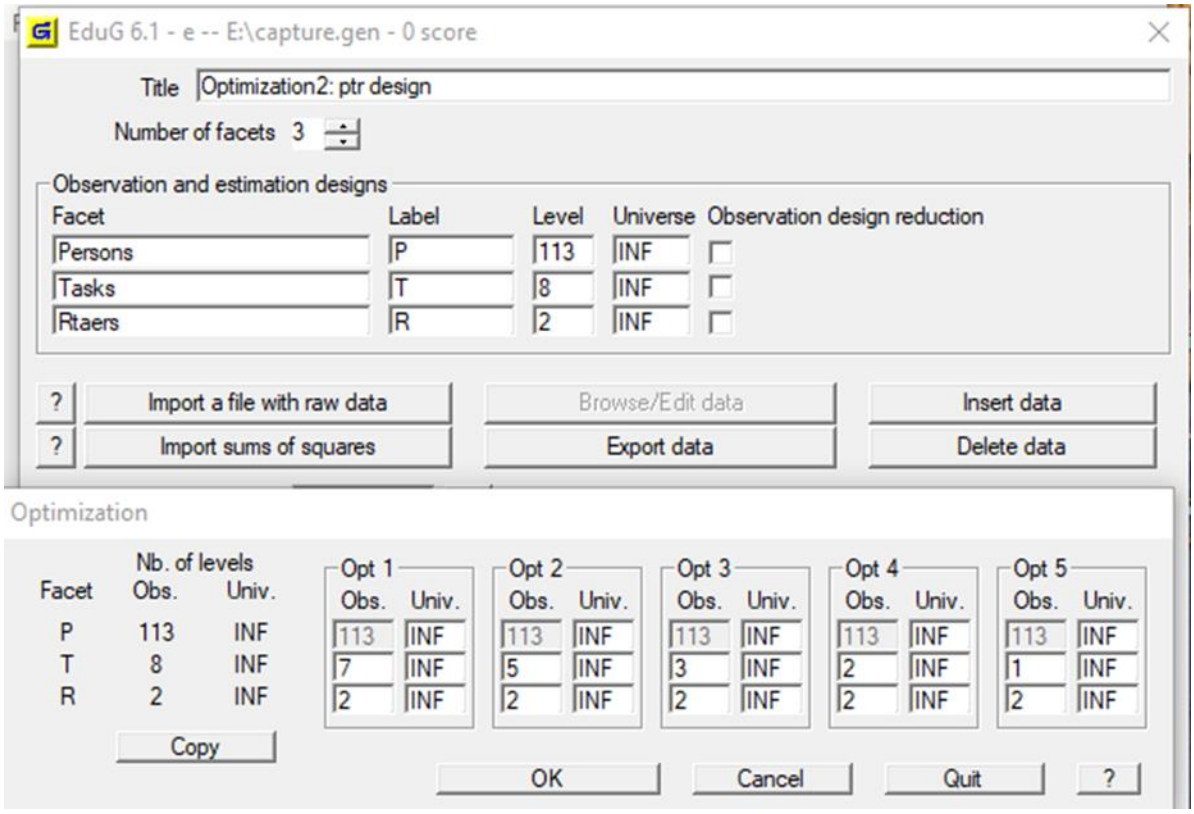


Decision studies (Optimization designs)

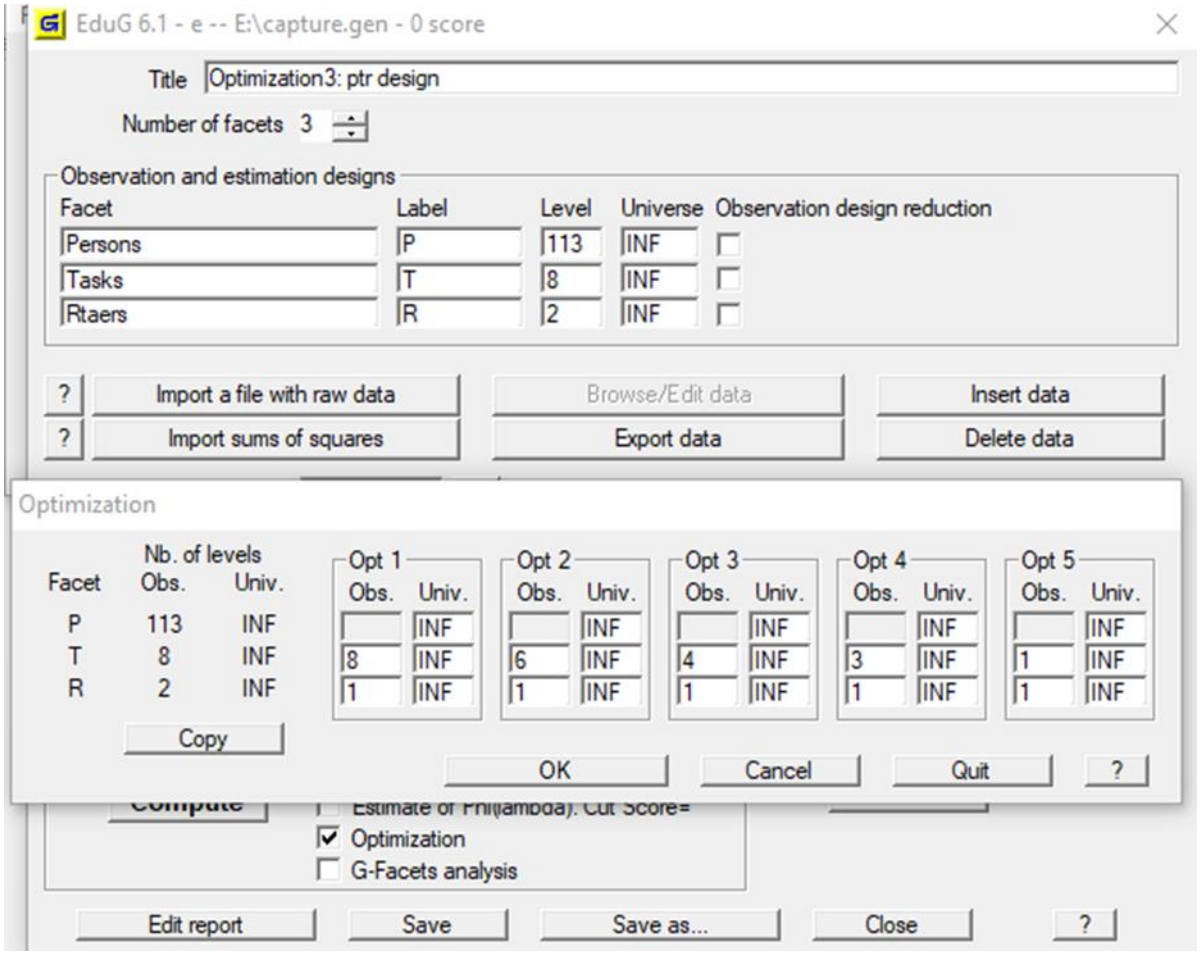
Optimization 1:



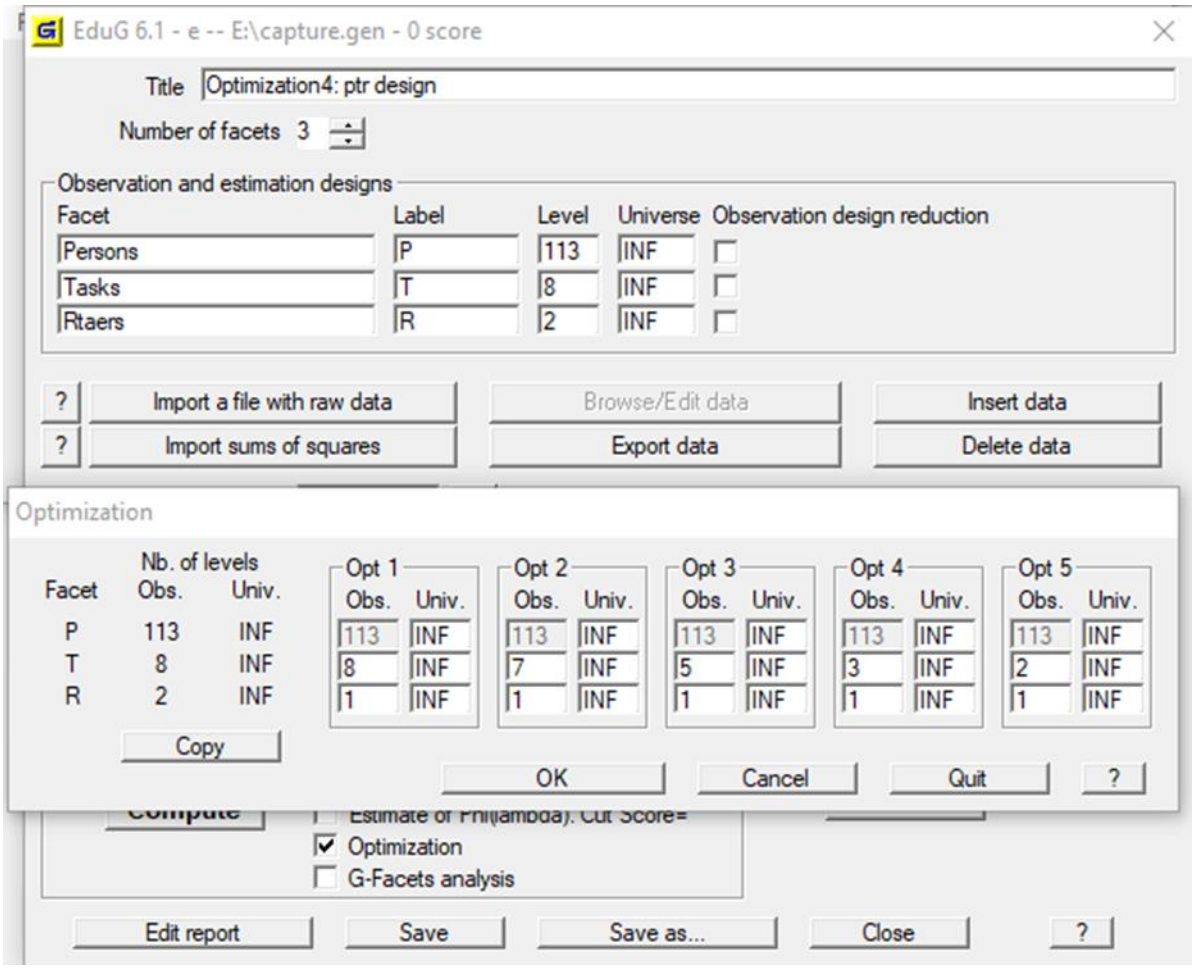
Optimization Design 2



Optimization Design 3:



Optimization 4:



الملخص

ينطوي "تقييم التقويم" على تحديد العوامل المؤثرة على جودته، والذي يبدو أنه بعيد عن الاهتمام إلى حدّ الآن. إن تطوير اجراءات التقويم التي تستهدف معرفة المفردات ومهارات استخدامها من قبل طلاب السنة الأولى جامعي في اللغة الانجليزية تحتاج إلى مزيد من التحسين بسبب مصادر خطأ القياس التي تهدد بشكل عام دقة واتساق التقويم. وهذا يؤكد على ضرورة تطبيق مبادئ نظرية إمكانية التعميم من أجل جمع أدلة عن الثبات والصدق التقاربي للدرجات الملاحظة في الاختبار. أجريت هذه الدراسة لتقدير الاتساق وعدم الاتساق في الدرجات المحصلة من الطلاب. تم استخدم المنهج الوصفي الاستكشافي لجمع البيانات الكمية من أجل تحديد مصادر الخطأ في الاختبار الذي أُجري على (113) طالبا مسجلا حديثا في قسم اللغة الانجليزية بـ "المدرسة العليا للأساتذة ببوزريعة" الذين أجابوا على اختبار المعرفة بالمفردات الانتاجية المعمقة. تم إعداد (08) مهام تواصلية تثير وضعيات معقدة لإظهار كفاءة الطلبة على تطبيق المعرفة بالمفردات المكتسبة مسبقا لحل مشكلات جديدة. تم تصحيح منتجات الطلبة بالاعتماد على مقيّميْن (02)، وتحليل البيانات باستعمال برمجية EduG من خلال تصميم (03) دراسات إمكانية التعميم، و(04) دراسات القرار. كان الهدف من دراسات القرار استخدام مصادر التباين المحددة في دراسات إمكانية التعميم من أجل تصميم أداة قياس يمكن أن تقلّل من حجم تباين الخطأ. كشفت دراسات إمكانية التعميم أن مصادر التباين التي أثرت بشكل كبير على ثبات الدرجات تُعزى إلى التفاعل بين الطالب والمهمة، والتفاعل بين الطالب والمقيم، ومكون الباقي. ومن جهة أخرى تأثر الصدق بتفاعل الطالب مع المهمة، وتفاعل الطالب مع الموضوع. أشارت دراسات القرار إلى أن معاملات إمكانية التعميم كانت مختلفة عبر تصميمات الدراسة، وأن الحد الأقصى لبلوغ مستويات مقبولة من إمكانية التعميم يتطلب (05) مهام ومقدّر واحد (01)، و(04) مهام ومقدريْن (02). أفضت هذه الدراسة إلى أن تقدير دقة القياس باستخدام نظرية إمكانية التعميم ذات أهمية بالغة للتحسين في نماذج التقويم والتصميمات.

الكلمات المفتاحية: تقييم الاداء؛ نظرية إمكانية التعميم؛ مصادر الخطأ أو التباين؛ الثبات؛ الصدق التقاربي؛ المهام التواصلية للمفردات؛ طلاب الإنجليزية كلغة اجنبية