ASIAN EFL JOURNAL
The EFL Professional's Written Forum

Construct Validity and Difficulty Index of Departmentalized Reading Comprehension Test for Grade 11 Students

Liane Neill N. Bautista
Pangasinan State University – School of Advanced Studies
Urdaneta City, Pangasinan Philippines

Presley V. De Vera, Ed. D., D. Comm.
Pangasinan State University
Lingayen, Pangasinan Philippines

Abstract

This is a descriptive study that assessed the construct validity of the departmentalized test for Grade 11 Literature implemented by PHINMA-University of Pangasinan in three (3) consecutive school years. In the context of this study, construct validity of the departmentalized test (DT) was investigated using twofold measures. The first one is carried out by assessing the DT's scope / coverage of learning competencies assessed. The second measure entails an evaluation of the appropriateness of the test types and questions of the DT in terms of their capability to assess students' "Knowledge, Understanding, and Mastery" of the literature course.

Findings revealed that the two parts of the DT (i.e. DT Part 1 and DT Part 2) fall short of their compliance in integrating all the learning competencies assigned to Grade 11 Literature course. Nevertheless, the overall DT's rate of compliance (65%) is "Very Satisfactory". On the bearing of these results to construct validity, the overall DT was found to have a "High level of Construct Validity" in terms of the evaluation of its scope / coverage of competencies

assessed. Moreover, the DT's overall rate of appropriateness is 4.40 (Highly Appropriate). This suggests that the types of test employed in the DT and the type of questions registered in the DT are assessed as "highly appropriate", generally speaking, in terms of how the test types and the questions contribute to the DT's capability to assess students' "Knowledge, Understanding, and Mastery" of the Grade 11 Literature course. The DT's difficulty index across different generations of students subjected to it has a consistent range of "Very Low" level of difficulty, thereby suggesting that the test's difficulty index is close to objective rather than context-sensitive.

Guidelines and action plan can be adopted to improve the construct validity of the subject DT, and these should be based on (a) the twofold assessment of the construct validity of the current DT used by the University and (b) the proposed compositional hierarchy of learning competencies assessed by the DT as perceived by teachers.

Based on the conclusions of the study, it recommends the need for the University to establish the complete set of learning competencies for the Grade 11 Literature subject, which defines what points should be assessed by the DT. Likewise, there is ample room to improve the adoption of more test types and questions to maximize the DT's capability to evaluate the students' knowledge, understanding, and mastery of the course. On the reports of the DT's difficulty index, it is recommended that prospective revisions of the DT should presuppose pre-testing in order to assess the feedback of test takers and treat them as inputs in the overall design of the DT. Teachers directly involved in the instruction of Grade 11 Literature subject must be consulted and directly involved in the deliberation and decisions as to what learning competencies are appropriate to be assessed by the DT. It is recommended that the University conducts continued monitoring of students' performance in the departmentalized test, as this may be treated as one basis to determine if the DT already requires revision at some point. Finally, the study modestly recommends the use of its proposed guidelines and action plan in

its pursuit to further improve the construct validity of its departmentalized test for Grade 11 Literature subject.

Keywords: *construct validity, difficulty index, departmentalized test*

Introduction

Background of the Study

Testing and evaluation of language skills and competencies are very important components of language teaching. In the particular context of this proposed study, it is specifically focused on testing and evaluation methods and instruments that are deployed in the academic teaching of English as a Second Language (ESL) which is integral to the curriculum of sophomore high school students as mandated by the Philippine Department of Education (DepEd), to which all Basic Education institutions in the country adhere to.

"Testing", even in most general sense, is considered as an integral part of teaching because it provides significant information or inputs about the growth and achievement of learner's difficulties, styles of learning, anxiety levels. Effective teaching and effective testing are two sides of the same coin. A curriculum is what constitutes a total teaching learning program composed of overall aims, syllabuses, materials, methods and testing in short. It provides a framework of knowledge and capabilities, selected to be appropriate to a particular level. Test evaluates not only the progress and achievement of learners but also the effectiveness of the teaching materials and methods used (Desheng & Verghese, 2013).

Asking students to demonstrate their understanding of a subject matter is critical to the learning process; it is essential to evaluate whether the educational goals and standards of the lessons are being met. In this context, it is clear that assessment is an integral part of instruction, as it determines whether or not the goals of education are being met. Assessment affects decisions about grades, placement, advancement, instructional needs, curriculum, and, in some

cases, funding. Assessment inspire us to ask these hard questions: "Are we teaching what we think we are teaching?" "Are students learning what they are supposed to be learning?" "Is there a way to teach the subject better, thereby promoting better learning?" (George Lucas Educational Foundation, 2018).

Today's students need to know not only the basic reading and arithmetic skills, but also skills that allowed them to face a world that is continually changing. They must be able to think critically, to analyze, and to make inferences. Changes in the skills base and knowledge our students need require new learning goals; these new learning goals change the relationship between assessment and instruction. Teachers need to take an active role in making decisions about the purpose of assessment and the content that is being assessed. (George Lucas Educational Foundation, 2018).

On the above explanations, the same level of priorities applies to the specific goals of academic English Language Teaching (ELT). In ELT, we are able to identify the difference among the macro skills of English, and which subset of micro skills register under them. Such identification becomes an instant reference as to what types of skills need to be assessed by the ELT teacher. However, as stipulated above, skills (whether these pertain to language or non-language skills) are imperative for students to be able to face a world that is continually changing. Hence, that we are able to primordially establish what skills they need to learn is not enough in such context. If skills are meant to address the challenges of a changing world, then so are the nature and demands of the skills also develop in complexity and sophistication to meet the ever advancing levels of challenges to which the skills apply to. Taking this in the concerns of language testing and evaluation, there is the risk that teachers may be setting an unchanging standard for the kind of skills they expect their students to develop. And this is reflected in the tests that teachers administer. A test always evaluates a learner on a particular basis and level of expectations. When expectations don’t change, then barely can there be

changes in the criteria and content of the tests as well. Tests provide the most succinct evidence if there are development in the contents and goals of an English language learning program.

The use of language testing, itself, promises a pool of significance and benefits for both the learners and the teacher. Tests provides “diagnostic feedback” (What is the student's knowledge base? What is the student's performance base? What are the student's needs? What has to be taught? It also helps educators “set standards” (What performance demonstrates understanding? What performance demonstrates knowledge? What performance demonstrates mastery?). Likewise, tests are used to “evaluate progress” (How is the student doing? What teaching methods or approaches are most effective? What changes or modifications to a lesson are needed to help the student? Relates to a student's progress What has the student learned? Can the student talk about the new knowledge? Can the student demonstrate and use the new skills in other projects?). Moreover, tests are also used to “motivate performance” on the part of the student and the teacher as well. In line with this, students may ask several questions (i.e. Now that I'm in charge of my learning, how am I doing? Now that I know how I'm doing, how can I do better? What else would I like to learn?). Teachers may also ask questions pertinent to their interests (What is working for the students? What can I do to help the students more? In what direction should we go next?). (George Lucas Educational Foundation, 2018).

Of the above categories of significance accorded to “testing and evaluation”, the researcher adopts the second category namely “the importance of tests in helping teachers set standards”. In the context of this category, there are three things that a language test, for that matter, must assume. It should be able to evaluate the student’s level of “knowledge”, level of “understanding”, and level of “mastery”. These become the major criteria for testing. These criteria actually informed the formulation of the proposed research questions of this study. The difference among “knowledge, understanding, and mastery” is better clarified when we link them to Desheng & Varghese’s (2013) “types of language tests”. The latter explained that:

> *"Language test broadly classified into two types as testing skills and testing knowledge of content. Skills such as listening, speaking, reading, and writing and sub- skills such as comprehension, vocabulary, grammar, spelling, punctuation, etc. Deferent kinds of tests are there to test student"s knowledge in language, the tests like non-referential test, aptitude test, proficiency test, achievement test and diagnostic test." (Desheng & Varghese, 2013)*

Reiterating from the above typology offered by Desheng & Varghese (2013), language tests are classified based on what exactly they intend to test, although, language tests are actually expected to presuppose that both types are exhaustively considered when preparing language tests. The two things that are generally evaluated by language tests are (a) language skills, and (b) knowledge and content (of formally instructed rules of the language). These types can be distributed to register the three testing criteria "knowledge, understanding, and mastery". For instance, "knowledge and understanding" register under "knowledge and content"; while "mastery" registers under "language skills". Hence, these are also suggestive of the two aspects of language testing namely "linguistic competence" and "linguistic performance", which are two polarized concepts (Chomsky, 1965). The aspect of a language test that deals on the assessment of linguistic competence is focused on "knowledge and content" or "knowledge and understanding". On the other hand, the aspect of the test that deals on linguistic performance aims to evaluate the students' "language skills" or "mastery". Anent these, herein proposed study delves into the evaluation of both aspects of a language test.

Another point emphasized in the above quote by Desheng& Varghese (2013), is that they also offer an alternative way of classifying language tests. Tests can also be classified according to which macro skill they intend to asses. To familiarize on what these different categories of language tests and evaluation are, which are differentiated from each other on the basis of the particular skills performance they aim to measure, Language Testing International or LTI (2018) provides an inventory, to wit:

> ***"Speaking Test.*** *A speaking proficiency test measures how fluently a*

person speaks when performing real-life communication tasks. Given that they will be the face of your company, you want to be sure that they are the best bilingual representatives possible.

Listening Test. *Testing the ability to understand what is being said to someone. Misunderstanding leads to frustration and dissatisfaction, and could prove detrimental to one's profits and future in a given community or country.*

Reading Test. *Testing the ability to read and understand a variety of informational texts, such as short messages, correspondence, and reports.*

Writing Test.*A writing test can be used to certify that employees are not only bilingual but biliterate, able to read and write in the target language". (Language Testing International, 2018)*

On the above classification, this study dealt with "reading test". However, even this type of language test comes in a very broad range because there are so many sub-skills and micro skills that register to "reading" as a macro skill. Among these, the focus here will be on reading skills applied to the text comprehension of literature, also known as "literature skills" (MobyMax, 2018). Accordingly, literature skills refer to "the specific skills of reading. Each reading skills lesson breaks the Common Core reading standards into small, achievable skills with targeted practice problems. It includes fundamental critical reading skills for both literature and information articles" (MobyMax, 2018).

The importance of an effective language test is indispensable to the interests and ends of students, the teachers and the school. On the part of students, the results of tests give them an impression about the progress of their learning, their strengths and weaknesses. On the part of teachers, the results of test enable them to have an overview of the strengths and weaknesses of their students along the different language competencies, which in turn, serve as their guide when choosing to reinforce instruction on certain competencies that seem to be more challenging or difficult for students. All these intended purposes of testing are barely achieved when the test themselves are not capable of producing adequate or accurate data on the reflection of the students' performance. In that way, results of test serve nothing more than the purpose of being a mere basis for the computation of student grades, but are not really useful and informative for purposes of improving literature pedagogy.

Different teachers of literature select their own focus in their respective delivery of instruction of a common subject matter. Teachers emphasize in their classrooms what they individually perceive to be the more exigent or important knowledge and skills that students ought to develop. Hence, it is not surprising when some teachers feel dismayed to later discover that the departmentalized tests administered to their students have only scarcely covered the competencies that they expected to be evaluated by the test. Moreover, some teachers may feel dismayed at the type of test used to evaluate their student's performance relative to certain competencies, or how certain test questions may have been poorly formulated to really reflect the genuine knowledge and skill of the students.

On these issues, the merits of employing "construct validity" to analyze language tests is set forth. In a way, the aim of this study is not just the mere conduct of an analysis over a particular departmentalized test using the framework of construct validity, but the further desire of the researcher to raise greater awareness about how important it is for language teachers and schools to check on the quality of the tests they employ. A lot of efforts are being spent to train teachers to keep them abreast with the most effective teaching strategies, and to keep them updated in the developments of the contents of the courses they teach. A lot of efforts are also spent for the selection of the most effective instructional materials or the provision of state-of-the-art teaching and learning instruments. All of these efforts can hardly be optimized of their benefits if, in the end, there is no effective mechanism for assessment that can tell if these intervention strategies really work as they are expected to work. One of the clearest basis for such an assessment is the result of tests that are administered on the students. And if the tests themselves are not effective, then it compromises, as well, the reliability of the data that can be obtain from the tests. This is the reason that this study was conducted, as it highlights the importance of efficiency and quality in the construction of language tests. For this study's limited scope, however, it focuses on the design and content of a departmentalized test intended

to asses Grade 11 students' performance in their corresponding literature course / subject, i.e. "21st Century Literature from the Philippines and the World".

This study provided a "description" as to which category of competency seem to receive greater emphasis in as far as the departmentalized test is concerned. In addition to this, it is also important to note that each of the three categories of competency presuppose the use of certain types of test that are designed to assess students' performance according to the differentiated ways of evaluating "knowledge, understanding, and mastery". On such basis, it is then important to know further how the departmentalized test accounts for such differentiation. Ultimately, the results of such analyses are synthesized as bases for drawing a framework of recommendations which this study proposed as a guide for improving the formulation of the departmentalized test.

Statement of the Problem

This study aimed to analyze the reading comprehension test integrated into the Departmentalized Tests administered to the Grade 11 students of PHINMA-University of Pangasinan. These tests refer to its current structure as implemented for the past 3 school year cycles until the present. The study utilized "construct validity" as the specific approach to determine the extent to which the test measure the ability that it is designed to assess. The construct validity analysis of the tests also investigated in relation to the test's index of difficulty based on the test performance of three batches of students across three school year cycles of its implementation. The merits of the findings served as bases for the recommendation of a framework for test construction and the concrete exemplification of this framework in a prototype test formulated by the researcher as the outcome of the study.

Specifically, it sought to answer the following questions:

1. What is the level of the construct validity of the departmentalized reading comprehension tests, in terms of:

a. scope / coverage of competencies assessed, and

b. appropriateness of the test type and questions in assessing the following categories of competencies on reading comprehension"

i. knowledge;

ii. understanding; and

iii. mastery?

2. What is the difficulty index of the departmentalized reading comprehension tests?

a. school year 2016-2017;

b. school year 2017-2018; and

c. school year 2018-2019

3. What are the perceptions of Literature teachers as to what learning competencies should be emphasized in the departmentalized reading comprehension tests for Grade 11 Literature subject?

4. Is there a significant difference among the difficulty index of the departmentalized reading comprehension tests during the school years 2016-2017, 2017-2018, and 2018-2019?

5. What guidelines can be developed to improve the construct validity of the departmentalized reading comprehension tests for Grade 11 Literature subject?

Related Literature

Language Testing

Language Testing is the practice and study of evaluating the proficiency of an individual in using a particular language effectively (Fulcher, n.d.). As a psychometric activity,

language testing traditionally was more concerned with the production, development and analysis of tests. Recent critical and ethical approaches to language testing have placed more emphasis on the uses of language tests. The purpose of a language test is to determine a person's knowledge and/or ability in the language and to discriminate that person's ability from that of others. Such ability may be of different kinds, achievement, proficiency or aptitude. Tests, unlike scales, consist of specified tasks through which language abilities are elicited. The term language assessment is used in free variation with language testing although it is also used somewhat more widely to include for example classroom testing for learning and institutional examinations (Fulcher, n.d.).

Construct Validity of Language Tests

Generally, construct validity is "the degree to which a test measures what it claims, or purports, to be measuring." (Brown, 1996). In the classical model of test validity, construct validity is one of three main types of validity evidence, alongside content validity and criterion validity (Guion, 1980). Modern validity theory defines construct validity as the overarching concern of validity research, subsuming all other types of validity evidence (Messick, 1995).

Construct validity is the appropriateness of inferences made on the basis of observations or measurements (often test scores), specifically whether a test measures the intended construct. Constructs are abstractions that are deliberately created by researchers in order to conceptualize the latent variable, which is correlated with scores on a given measure (although it is not directly observable). Construct validity examines the question: Does the measure behave like the theory says a measure of that construct should behave? (Wikipedia: on "Construct Validity")

Construct validity is essential to the perceived overall validity of the test. Construct validity is particularly important in the social sciences, psychology, psychometrics and

language studies. Psychologists such as Samuel Messick (1998) have pushed for a unified view of construct validity "...as 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...".

Key to construct validity are the theoretical ideas behind the trait under consideration, i.e. the concepts that organize how aspects of personality, intelligence, etc. are viewed. Paul Meehl states that, "The best construct is the one around which we can build the greatest number of inferences, in the most direct fashion." (Wikipedia: on “Construct Validity”). Scale purification, i.e. "the process of eliminating items from multi-item scales" (Wieland, 2017) can influence construct validity. A framework presented by Wieland (2017) highlights that both statistical and judgmental criteria need to be taken under consideration when making scale purification decision.

Research Methodology

This study adopted a descriptive design. Descriptive research, from a general perspective, is concerned with the conditions or relationships that exist, opinion that hold processes, effects that are evident, trends that are developing and describes the data and characteristics about the population or phenomena being studied (Bhat, 2018). Bhat further specifies that descriptive research engages in several types of investigations, which include (a) defining respondent characteristics; (b) measuring data trends; (c) conducting comparison; and (d) validating existing conditions; and (e) conducting research at different times. As regards these, some of the investigative procedures that are adopted by this study involve any one or a combination of the aforementioned items. Instead of profiling respondents, a Departmentalized Language Test was profiled along several criteria. Data trends are obtained from a comparative analysis of the difficulty index of the Departmentalized Test across three years of its

implementation, and with three different batches of students to whom it was administered.

Research Instrument

The conduct of this study presupposed the use of several instruments intended to gather the different sets of data as required in answering each of the research questions. The following discusses each of the proposed instruments.

Construct Validity Questionnaire (CVQ)

The idea of a construct validity questionnaire is based on the article "Construct Validity" published by Lund Research (2012). The article quotes:

> *"[…] construct validity can be viewed as an overarching term to assess the validity of the measurement procedure (e.g., a questionnaire) that you use to measure a given construct. This is because it incorporates a number of other forms of validity (i.e., content validity, convergent and divergent validity, and criterion validity) that help in the assessment of such construct validity" (Messick, 1980 in Lund Research, 2012).*

The purpose of this instrument is to generate a qualitative analysis of the Departmentalized Language Test for Grade 11 English used by PHINMA University of Pangasinan. Analysis will focus on the following:

a. scope / coverage of competencies assessed, and
b. appropriateness of the test type and questions in assessing the following categories of competencies on reading comprehension
    i. knowledge;
    ii. understanding; and
    iii. mastery

However, unlike other questionnaires used in research which were administered to respondents, the CVQ only served as a guide for this researcher to formally analyze the construct validity of the subject Departmentalized Language Test. The questionnaire items are also assigned a point system allowing for a quantitative transmutation of the results of the analysis that can serve as basis for computing the so-called rate of construct validity. The items of the CVQ are based on criteria appropriate for construct validity as discussed in the related

literatures that have been reviewed by the researcher, including criteria that are based on the curriculum guide for English 11 used by PHINMA-University of Pangasinan. A copy of this curriculum guide is indicated in Appendix A. A copy of the CVQ instrument is shown in Appendix B. The scoring rubric and formula for obtaining the construct validity rate is found in the final section of the instrument.

Language Test Diagnosis Perceptual Questionnaire (LTDPQ)

The purpose of this questionnaire is to obtain data from the perception of teachers in charge of Grade 11 English as to their assessment of the Departmentalized Test in terms of what it should cover or contain. Data generated from this questionnaire served as an additional input in the development of the guidelines for constructing a language test, which is the intended output of the study, was formulated in consonance with the guidelines. A copy of this questionnaire is found in Appendix C.

Data Gathering Procedure

Prior to administering the data-gathering instruments, permission was officially sought from the Administration, and all concerned offices of the PHINMA-UPANG College Urdaneta City. For transparency purposes, the details, analytical procedures and the ethical clause of the study was properly informed through either or both written and oral communication during the negotiations. The sample communication for the aforementioned concern is attached in Appendix D.

On the assumption that all negotiations have been established, and that permission to conduct the study and data retrievable have been granted by the University Administration, the researcher proceeded to retrieve the copies of the departmentalized test for Grade 11, in its versions used during school year 2016-2017, school year 2017-2018, and school year 2018-2019. Likewise, the researcher also retrieved from the concerned office the actual copies of the

test papers filled out by the students which have already been previously checked. A total of thirty (30) test papers were taken to represent each of the three school years (SY) when that the test was been administered. From the 30 test papers, 15 should come from male test-takers and 15 from the female test-takers. With a total of 30 tests per school year, the grand total of test papers that were retrieved for the three school year is 90. Copies of the test papers were handled with utmost care and confidentiality and were duly returned to the concerned office, once the tabulation of the scores have been obtained. Following this procedure, the Language Test Diagnosis Perceptual Questionnaire were ready to be administered to the teacher-respondents. They took home this questionnaire to allow them ample time to indicate their responses, and the filled-out forms were collected from them after three (3) days.

Results and Discussions

Scope / coverage of competencies assessed

This subsection presents the findings of the DT's construct validity through the use of the first construct validity measure. In this measure, the objective is to determine the extent of the scope / coverage of the competencies assessed by the DT compared against the competency standards prescribed by PHINMA-University of Pangasinan in the curriculum of the Grade 11 Literature course. Table 1 shows the findings.

First thing to note about Table 1, the latter shows the inventory of the course competencies placed under the first column. Moreover, the competencies are divided into two periodic terms. Likewise, it was also observed and noted that the University's curriculum guide (CG) for the Grade 11 Literature course does not provide distinction between course topics and competencies so that the course topics themselves are reflected to indicate the specific course competencies.

Table 1

Scope / Coverage of Competencies Assessed

First Quarter / Periodic Term: Scope / Coverage of Competencies *(with*	**%**	**Compliance Status**	**Test Items (No. of test items)**	**Rate of Compliance*** **
Defining Literature	20%	Null (0%)	N/A	80% (Very Satisfactory)
Describing literature in the pre-colonial times.	20%	Complied (20%)	Test I: 1, 2, 3 (3)	
Riddles, Salawikain, Short Poems and Songs Epics and Myths Folktales, Writing Folktale	20%	Complied (20%)	Test I: 4, 5 Test II: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 (16)	
Appreciating the contributions of the canonical Filipino writers to the development of national literature.	20%	Complied (20%)	Test II: 15 Test III: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (11)	
Describing Literature under Spanish Colonialism Describing Literature under American Colonialism Describing Philippine Literature under the Republic Describing Philippine Literature after EDSA Describing Philippine Literature after EDSA	20%	Complied (20%)	Test II: 11, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 (11)	
Second Quarter / Periodic Term: Scope / Coverage of Competencies				
Remembering Martial Law Poverty Gender Inequalities Justice System The Philippine Diaspora The Philippine Diaspora The Philippine Diaspora	50%	Complied (50%)	Test I: 1, 2, 11, 12 Test III: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (14)	50% (Satisfactory)

Fantasy, Horror, and the Old Country				
Identifying the Roles of Multi-Media in Literature Producing a creative representation of a literary text by applying multi-media skills Digital Presentation of a Literary Piece	50%	Null (0%)		
				65% (Very Satisfactory)

Legend: % --- Percent of allocation in the overall Quarter Curriculum

*** Rubric on the level of Construct Validity based on the D|T's rate of compliance indicative of its scope / coverage of competencies assessed

Range (in terms of percentage)	**Description (rate of compliance)**	**Interpretation (level of construct validity)**
100 – 81	Outstanding	Very High
80 – 61	Very Satisfactory	High
60 – 41	Satisfactory	Moderate
40 – 21	Fair	Low
20 – 0	Poor	Very Low

Appropriateness of the Test Type and Questions in Assessing Knowledge, Understanding; and Mastery

This subsection presents the findings of the DT's construct validity through the use of the second construct validity measure. In this measure, the objective is to determine the extent of appropriateness of the test types and questions used in the DT as they fare to assess "knowledge, understanding, and mastery" of the topics and competencies assigned by PHINMA-UPANG for the curriculum of Grade 11 Literature course. Table 2 shows the findings.

The structure of Table 2 indicates a clustering of the construct validity indicators into three groups labeled as "Knowledge, Understanding, and Mastery". The table reflects the ratings assigned to each indicator, ranged from 1 (lowest) to 5 (highest). Moreover, the "mean"

of the ratings per indicator is also indicated under the seventh column, with their corresponding "description" in terms of "rate of appropriateness", ranging from "High" (highest) to "Very Poor" (lowest) rates of appropriateness of the DT's test type and questions. A summary table is provided underneath the larger table bearing a synthesis based on the raw data.

A careful assessment of the DT reveals its overall rate of appropriateness at 4.40 (Highly Appropriate). This means that the types of test employed in the DT and the type of questions registered in the DT are judged to "highly appropriate", generally speaking, in terms of how the test types and the questions contribute to the ST's capability to assess "Knowledge, Understanding, and Mastery" of the Grade 11 Literature course. Further interpreted in terms of test construct validity, the DT has a construct validity level of "High", which represents the highest scale of test construct validity relative to the rubric used in this study. However, a more detailed look into this assessment result shows the variables that contribute to this overall rating of the DT. These details are explained as follows.

In terms of the DT's capability to assess the Grade 11 students' "Understanding" and "Mastery" of the Literature course, the DT received ratings of 4.62 (Highly Appropriate) and 4.34 (Highly Appropriate) respectively for the two indicators mentioned.

Table 2

Appropriateness of Departmentalized Test Type and Questions in terms of the Indicators of Construct Validity (Knowledge, Understanding, Mastery)

Indicators	1	2	3	4	5	Mean	Description
Knowledge							
The test evaluates students' ability to answer questions that require objective answers (what, who, where, when, which one)	1	0	4	6	9	4.10	MA
Test questions requiring objective answers are carefully formulated to eliminate ambiguity that	0	0	2	8	10	4.40	HA

misleads the student to a different answer							
Test questions requiring objective answers are carefully formulated to eliminate the possibility of having more than one possible answer	0	0	1	10	9	4.40	HA
Test questions requiring objective answers are carefully formulated to eliminate difficult words or sentence structures that interferes with the clarity of the questions	0	0	1	10	9	4.40	HA
Most of the questions of the test require objective answers only (what, who, where, when, which one)	1	3	5	7	4	3.50	**MA**
					Category Average	**4.16**	**MA**
Understanding							
The test includes questions that require higher level thinking aside from questions merely requiring objective answers	0	0	0	5	15	4.75	HA
the test questions require students to engage in inferential, and creative levels of comprehension (why, how, what is the implication of, etc.)	0	0	2	2	16	4.70	HA
The test questions do not necessarily reflect the same explanations given to concepts as they appear in books or instructional materials to allow students to really understand the question	0	0	1	7	12	4.55	HA
Test questions requiring "understanding" are carefully formulated to eliminate ambiguity that misleads the student to a different answer	0	0	1	4	15	4.70	HA
Test questions requiring "understanding" are carefully formulated to eliminate the possibility of having more than one possible answer	0	0	0	7	13	4.65	HA
Test questions requiring "understanding" are carefully formulated to eliminate difficult words or sentence structures that	0	0	0	8	12	4.60	HA

interferes with the clarity of the questions							
Most of the questions of the test require "understanding"	0	1	1	7	10	4.37	**HA**
						Category Average	**4.62**	**HA**
Mastery							
The test features questions that laden with complexity that are challenging to students	0	0	0	10	10	4.50	HA
The test features questions that are beyond what has been taken up by the course but can be answered based on what has been studied in the course	0	0	4	8	8	4.20	MA
The test incorporates questions that are seemingly repetitive but structured differently for purposes of testing the consistency of students' answers	0	0	3	10	7	4.20	MA
Test questions requiring "mastery" are carefully formulated to eliminate difficult words or sentence structures that interferes with the clarity of the questions	0	0	2	7	11	4.45	HA
						Category Average	**4.34**	**HA**
Overall Average							**4.40**	**HA**

Legend: M\A – Moderately Appropriate (3.41 – 4.20), HI – Highly Appropriate (4.21 – 5.00)

Construct Validity Indicators	Rating (AWM)	Description (rate of appropriateness)	Interpretation (Level of Construct validity)
Knowledge	4.16	Moderately Appropriate	Moderate
Understanding	4.62	Highly Appropriate	High
Mastery	4.34	Highly Appropriate	High
Synthesis	4.40	Highly Appropriate	High

DIFFICULTY INDEX OF THE DEPARTMENTALIZED TEST ACROSS A THREE-SCHOOL YEAR IMPLEMENTATION PERIOD

Table 3

Report of Difficulty Index of the Departmentalized Test across Three School Years***

Departmentalized Test	School Year	Number of Difficult Items	Difficulty Index (% of identified difficult items)	Description (Level of Test Difficulty)
Part 1 (First Quarter / Grading Period)	SY 2016-2017	0	0%	Very Low
	SY 2017-2018	1	2.5%	Very Low
	SY 2018-2019	0	0%	Very Low
Part 2 (First Quarter / Grading Period)	SY 2016-2017	0	0%	Very Low
	SY 2017-2018	9	18%	Very Low
	SY 2018-2019	0	0%	Very Low

*** Data entries in Table 3 are based on the reference tables in Appendix H.1 and H.2 that show the results of the item analysis of the Departmentalized Test across three (3) school years

Based on the above-mentioned findings, a clear perceivable pattern can be established on the difficulty index of the DT as a whole. First, it is noteworthy that the two parts of the DT (i.e. DT-Part 1 and DT-Part 2) consistently registered a "Very Low" difficulty level. And this is true across the three school years. Overall, the DT registers a "Very Low" level of difficulty. As much as the difficulty index has been tested across three generations of Grade 11 students, the consistency in the findings reinforce its reliability and generalizability.

Table 4 shows that the most number of learning competencies registered under "Average Importance" in the compositional hierarch (6 or 46%). A lesser number of the learning competencies (5 or 39%) registered under "Low Importance". Surprisingly, the least number of learning competencies (2 or 15%) registered under "Prime Importance".

First thing suggested by this compositional hierarchy of the DT is that teachers find only very few of the learning competencies that should be emphasized by the DT. There can be different ways to interpret this finding. It may be suggestive of a fact that the teachers find only a few of the prescribed learning competencies to have prime importance. Or, the findings may also be interpreted another way that the teachers only find a few of the learning competencies testable based on the format or objective of the departmentalized test, i.e. however DT is assigned a significance in the assessment of students' performance. The value of a DT is usually arbitrary and is customized by an institution (Chennis, 2018). Likewise, the test format (i.e. type/s of test) featured in the DT is institutionally arbitrary and customized. It was beyond the framework of this study to make further clarification as to why only a few learning competencies are ranked with prime importance. It is recommendable for future research to include this aspect in the research framework.

COMPOSITIONAL HEIRARCHY OF LEARNING COMPETENCIES ASSESSED BY THE DEPARTMENTALIZED TEST FOR GRADE 11 LITERATURE SUBJECT AS PERCEIVED BY TEACHERS

Table 4

Compositional Hierarchy of Learning Competencies assessed by the Departmentalized Test
n = 20

Indicators	Number / Rate	Mean Rank	Hierarchy Category
The test assesses the way students appreciate the literature produced in other regions of the Philippines	2 / 15%	2.69	PI
The test enables the students to determine the impact of historical, political, social, or economic developments of the Philippines in the 21st Century to the contents of literary pieces produced locally and by other regions of the country		3.30	PI
The test assesses the way students appreciate the literature produced in the locale (where the school is located)		3.73	AI

The test enables the students to draw the implications of literary pieces based on the historical, political, social, or economic developments of the Philippines in the 21st Century	6 / 46%	3.75	AI
The test assesses the way students compare the literature produced in the locale (where the school is located) and those from other regions of the country		4.59	AI
The test evaluates students' ability to compare and contrast the features of literature produced in different times or eras.		5.19	AI
The test includes evaluation of students' familiarity with literature generated from different periods (e.g. Spanish Colonialism; American Colonialism; the Republic; after EDSA)		6.38	AI
The test evaluates students' ability to identify the elements and features of various literary genres (e.g. Riddles, Salawikain, Short Poems and Songs, Epics and Myths, Folktales, Writing Folktale		6.65	AI
The test evaluates students' knowledge about the contributions of the canonical Filipino writers to the development of national literature.	5 / 39%	6.86	LI
The test evaluates' students' knowledge about literature on various topics (please scale the items below:)		6.90	LI
The test allows students to identify the Roles of Multi-Media in Literature		7.33	LI
The test enable students to produce creative representation of a literary text by applying multi-media skills		7.85	LI
The test evaluates students' familiarity in generating Digital Presentation of a Literary Piece		9.00	LI

Legend:

	Prime Importance (PI)	1.0- 3.33	Must be substantiated by the items of the DT Must be given foremost allocation in the DT
	Average Importance (AI)	3.34 – 6.66	Must be substantiated by the items of the DT Given fair allocation in the DT
	Low Importance (LI)	6.67 - 10	Substantiated by the items of the DT, although negligible May or may not be given allocation in the DT

DIFFERENCE IN DIFFICULTY INDEX OF THE DEPARTMENTALIZED TEST ACROSS A THREE-SCHOOL YEAR IMPLEMENTATION PERIOD

Table 5 shows the results of the statistical computation using "Analysis of Variance" (ANOVA) in determining any significance in the difference of the DT's difficulty index reports respective to the three school years to which the test's difficulty index was obtained. Repeated measures analysis of variance was employed to determine if difficulty indices are significantly different among the three academic years. Doing so, the Greenhouse-Geisser statistic is found to be 61.264 with an associated significance value equal to 0.000. These values imply that there is a significant difference in the difficulty indices among the three concerned school years.

Looking into which school years differ, using Schefft's post-hoc analysis, it was found out that school years 2016 – 2017 and 2018 – 2019 are similar. This means that they are not significantly different. However, school year 2017 – 2018 is found to be significantly different from the other school years.

Table 5

Analysis of Variance (ANOVA) of the Difficulty Index Reports of the Departmentalized Tests across Three School Years

Source		F	Sig.	School Year	Mean	Grouping
Academic Year	Greenhouse-Geisser	61.264	0.000**	2016 - 2017	87.21	A
				2017 - 2018	71.82	B
				2018 - 2019	93.27	A

** - Significant at 1% level of significance.
Academic years with different Grouping values are significantly different.

The last paragraph's implication is evident in the means of the difficulty indices. The difficulty index for school year 2017 – 2018 can be said to be much lower than those of 2016 – 2017 and 2018 – 2019.

Conclusions

Based on the merits of the findings, the following conclusions are drawn:

1. The overall departmentalized test has "high level of construct validity" in terms of the test's scope / coverage of competencies assessed. Parallel to this, the departmentalized test also obtained a rating of "high level of construct validity" in terms of the appropriateness of its test type (format) and questions in assessing students' knowledge, understanding, and mastery of the Grade 11 Literature course.

2. The departmentalized test's difficulty index ranges from 0 (minimum) to 1 (maximum).

3. As perceived by the English teachers, the learning competencies for Grade 11 Literature subject ranked differently in terms of their importance as objects to be assessed in the departmentalized test.

4. The difficulty index of the departmentalized test is context-sensitive as it significantly differs based on the generation of students who are subjected to the test.

5. To improve the construct validity of the departmentalized test for Grade 11 Literature subject, guidelines and an action plan can be designed based on the assessment of the test's scope / coverage of learning competencies assessed, appropriateness of the test's format and questions in assessing knowledge, understanding and mastery of the course, and the compositional hierarchy of learning competencies that the test aims to assess.

Reference

Books

Brown, J.D. (1996). **Testing in Language Programs**. NJ: Prentice Hall.

Chomsky, N. (1965). **Aspects of the Theory of Syntax**. Cambridge, MA: MIT Press.

Shields, P. & N. Rangarajan (2013). **A Playbook for Research Methods: Integrating Conceptual Frameworks and Project Management**. Oklahoma: New Forums Press.

Unpublished Dissertation

Brown, J.D., C. Chaudron, T.D. Hudson, G. Kasper, and P. Chandler (2004). **Validity Evaluation in Foreign Language Assessment**. Ph.D. Dissertation: University of Hawaii. PDF file.K to 12 Senior High School Core Curriculum (2013 ed.). 21st Century Literature from the Philippines and the World (Grade 11 / 12).

Electronic Sources

Bhat, A. (2018). Descriptive Research: Definition, Characteristics, Methods, Examples and Advantages. QuestionPro. Retrieved from https://www.questionpro.com/blog/descriptive-research/.

Brown, J.D. (2000). What is Construct Validity? JALT. Retrieved from http://hosted.jalt.org/test/bro_8.htm.

Center for Teaching Excellence (2018). Preparing Tests and Exams. University of Waterloo. Retrieved from https://uwaterloo.ca/centre-for-teaching-excellence/teaching-resources/teaching-tips/developing-assignments/exams/exam-preparation.

Chennis, St.T. (2018). The Impact of Traditional and Departmentalized Classroom Instructional Settings on Fifth Grade Students' Reading. Liberty University. Retrieved from https://digitalcommons.liberty.edu/cgi/viewcontent.cgi?article=2806&context=doctoral.

Chestnut, D. (2018). How to Calculate Difficulty Index? The Classroom. Retrieved from https://www.theclassroom.com/calculate-difficulty-index-8247462.html.

CIIT College of Arts & Technology (2017). Steps to Knowing which Senior High School Track Best Fits You. Retrieved from https://www.ciit.edu.ph/senior-high-school-track/.

Cole, N.L. (2018). Understanding Descriptive vs. Inferential Statistics. ThoughtCo. Retrieved from https://www.thoughtco.com/understanding-descriptive-vs-inferential-statistics-3026698.

Collins, R. (2014). Skills for the 21st Century: Teaching Higher-Order Thinking. Curriculum and Leadership Journal, 12(14). Retrieved from http://www.curriculum.edu.au/leader/teaching_higher_order_thinking,37431.html?issueID=12910.

Coughlin, M. (2019). Creating a Quality Language Test. UsingEnglish.Com.

Retrieved from https://www.usingenglish.com/articles/creating-quality-language-test.html.

Desheng, C. & A. Verghese (2013). Testing and Evaluation of Language Skills. IOSR Journal of Research & Method in Education, 1(2),31-33. Retrieved from http://www.iosrjournals.org/iosr-jrme/papers/Vol-1%20Issue 2/F0123133.pdf?id=1662.

DeWitt, P. (2015). Does Subject-matter Knowledge Count as much as We Think?

Education Week. Retrieved from http://blogs.edweek.org/edweek/finding_common_ground/2015/09/does_subject-matter_knowledge_matter_as_much_as_we_think.html.

Fulcher, G. (n.d.). What is Language Testing? Retrieved from http://languagetesting.info/whatis/lt.html.

George Lucas Educational Foundation (2018). Why is assessment important? Edutopia. Retrieved from https://www.edutopia.org/assessment-guide-importance.

Gosselin, D. (2017). Competencies and Learning Outcomes. InTeGrate. Retrieved from https://serc.carleton.edu/integrate/programs/workforceprep/competencies_and_LO.html.

Guion, R.M. (1980). On Trinitarian Doctrines of Validity. Professional Psychology, 11(3),385-398. Retrieved from http://psycnet.apa.org/record/1981-22475-001.

Hakuta, K. & L.L. Jacks (2009). Guidelines for the Assessment of English Language Learners. Educational Testing Service. Retrieved from https://www.ets.org/s/about/pdf/ell_guidelines.pdf.

Koksal, D. & K. Cesur (2012). Students and Instructors' Perceptions of Objective Tests Used to Assess Language Performance at University Level.Academia. Retrieved from https://www.academia.edu/37794950/STUDENTS_AND_INSTRUCTORS_PERCEPTIONS_OF_OBJECTIVE_TESTS_USED_TO_ASSESS_LANGUAGE_PERFORMANCE_AT_UNIVERSITY_LEVEL.

Language Testing International (2018). Learn about language testing and assessment. Retrieved from https://www.languagetesting.com/language-testing-and-assessment.

Larsson, J. & I. Holmstrom (2007). Phenomenographic or phenomenological analysis: Does it matter? International Journal of Qualitative Studies on Health and Well-Being. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/17482620601068105.

Lazaraton, A. & L. Taylor (2007). Qualitative Research Methods in Language Test Development and Validation. University of Ottawa Press. Retrieved from https://books.openedition.org/uop/1570?lang=en.

Lopez, A. (2010). Validation Study of Colombia's ECAES English Exam. ResearchGate. Retrieved from

https://www.researchgate.net/publication/319768768_Validation_Study_of_Colombia's_ECAES_English_Exam.

Lund Research (2012). Construct Validity. Laerd Dissertation. Retrieved from http://dissertation.laerd.com/construct-validity.php.

Macatangay, N. (2014). Departmentalized Examination. Prezi. Retrieved from https://prezi.com/pw_cu_oktove/departmentalized-examination/.

McMillan, J.H. & J. Hearn. 2008. Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement. Retrieved from https://files.eric.ed.gov/fulltext/EJ815370.pdf.

McNamara, T. (2010). The use of language tests in the service of policy: issues of validity. Retrieved from https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2010-1-page-7.htm.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9),741-749. Retrieved from http://psycnet.apa.org/doiLanding?doi=10.1037%2F0003-066X.50.9.741.

Messick, S. (1995). Test Validity: A Matter of Consequence. Social Indicators Research,45(1-3),35-44. Retrieved from https://link.springer.com/article/10.1023%2FA%3A1006964925094.

MobyMax (2018). Reading Skills Literature. Retrieved from https://www.mobymax.com/curriculum/reading-skills-literature.

Mohammad Ali, C. and R. Sultana (2016). A Study of the Validity of English Language Testing at the Higher Secondary Level in Bangladesh. International Journal of Applied Linguistics & English Literature, 5(6). Retrieved from http://www.journals.aiac.org.au/index.php/IJALEL/article/view/2599.

Oller, J., K. Perkins, F. Butler, and K. Krug (1980). Research in Language Testing. Research Gate. Retrieved from https://www.researchgate.net/publication/308555814_Research_in_Language_Testing.

Ozera, I. S.M. Fitzgeralda, E. Sulbarana, and D. Garveya (2013). Reliability and content validity of an English as a Foreign Language (EFL) grade-level test for Turkish primary grade students. Procedia. Retrieved from https://ac.els-cdn.com/S1877042814012671/1-s2.0-S1877042814012671-main.pdf?_tid=667be726-b5f6-48f8-aad6 1dc1881eb65b&acdnat=1544871007_b9a24ea96032ee5abb9ed1eadf37bbd9.

Powers, D.E. (2010). The Case for a Comprehensive, Four-Skills Assessment of English Language Proficiency. TOEIC. Retrieved from https://www.ets.org/Media/Research/pdf/TC-10-12.pdf.

Quileste, R. (2015). Item Analysis. All You Need to Know About It. Linkedin. Retrieved from https://www.slideshare.net/RonaldQuileste/item-analysis-discrimination-and-difficulty-index.

Renner, R. (2019). How to Calculate Difficulty Index. The Classroom. Retrieved from https://www.theclassroom.com/calculate-difficulty-index-8247462.html.

Siddiek, A.G. (2010). The Impact of Test Content Validity on Language Teaching and Learning. Research Gate. Retrieved from https://www.researchgate.net/publication/47807762_The_Impact_of_Test_Content_Validity_on_Language_Teaching_and_Learning. Retrieved: 15 December 2018.

Strauss, M.E. and G.T. Smith (2009). Construct Validity: Advances in Theory and Methodology. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2739261/.

Wieland, A. (2017). Statistical and Judgmental Criteria for Scale Purification. EmeraldInsight. Retrieved from https://www.emeraldinsight.com/doi/full/10.1108/SCM-07-2016-0230.

Wikipedia: "Compositional Containment Hierarchy" (subset of the topic "Hierarchy"). Retrieved from https://en.wikipedia.org/wiki/Hierarchy.

Wikipedia. "Construct Validity". Retrieved from https://en.wikipedia.org/wiki/Construct_validity.