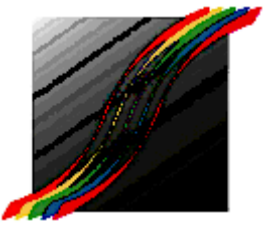
ISSN 20194-5876

Educational Measurement and Evaluation Review

JULY 2018

THE PHILIPPINE EDUCATIONAL MEASUREMENT AND EVALUATION ASSOCIATION (PEMEA)

The Educational Measurement and Evaluation Review (EMEReview) is the official publication of the Philippine Educational Measurement and Evaluation Association (PEMEA). It is international, refereed, and abstracted/indexed. The EMEReview publishes scholarly reports about contemporary theories and practices in the field of education and social science that highlights measurement, assessment, and evaluation. It welcomes articles that are about test and scale development, quantitative models of a construct, evaluation studies, best practices in evaluation, issues and policies on assessment, contemporary approaches in educational and psychological measurement, and other studies with direct implication to assessment in education, social science, and related fields. EMEReview is indexed/abstracted in the Open J-Gate, JournalTOCs, Google Scholar, InfoBase Index, Social Science Research Network, Open Academic Journals Index, Scientific Indexing Services, and ejournals.ph



Publication Division of PEMEA
Philippine Educational Measurement and Evaluation Association

Educational Measurement and Evaluation Review
ISSN 2094-5876 http://www.pemea.org/emereview
◈ July 2018, Vol. 9 Issue 1

Exploring Filipino Teachers' Conceptions of Assessment

Christine Joy A. Ballada and Marie Antoniette C. Alino 1

Development and Validation of Pagbabaybay (Spelling) and Pagkilala sa Salita (Word Recognition) of the Filipino Reading Achievement Test

Ryan Francis O. Cayubit, Lyka Ilonah D.C. Chua, Emerald Ann S. David, Therese Monique D.G. Gutierrez, Shiara Marriz T. Marquez, Niko A. Mendoza, Emille Joyce P. Palogan, and Reniel B. Tiu 24

Gender Differential Item Functioning in Polytomous Items: A Comparison of Three Methods

Consuelo T. Chua, Jose Q. Pedrajita, and Kevin Carl P. Santos 45

Moderating Role of Defensive Pessimism in the Relationship Between Test Anxiety and Performance in a Licensure Examination

Rene M. Nob, Alyonna Marie L. Bumanglag Genevie Mae A. Diwa, and Guia Isabel Ponce 67

Assessing the Construct Validity of the Locus-of-Hope Scale

Dominique T. Rivera and Leny G. Gadiana 84

Educational Measurement and Evaluation Review (2018), Vol. 9 Issue 1

Exploring Filipino Teachers' Conceptions of Assessment

Christine Joy A. Ballada
De La Salle University, Manila

Marie Antoniette C. Aliño
St. Paul University-Quezon City

Abstract

Teachers' knowledge and beliefs about the teaching and learning process influence their classroom practices. One important component of this teaching-learning process that is affected by teachers' beliefs is classroom assessment. This study sought to examine the structure of Filipino teachers' beliefs, meanings, propositions, rules, and mental images, more generally referred to as conceptions about assessment. A total of 391 Filipino teachers responded to the Conceptions of Assessment Inventory (COA-III; Brown, 2004). A combination of confirmatory and exploratory factor analyses revealed that the structure of Filipino teachers' conceptions of assessment may be explained by three dimensions: (1) assessment as a means to improve teaching and learning; (2) assessment as a means to hold schools accountable; and (3) assessment as irrelevant to teaching and learning. Implications for pre-service and in-service teacher training and further research are discussed.

Keywords: Conceptions of assessment, beliefs about assessment, Filipino teachers

Introduction

It is generally accepted that teachers' knowledge, beliefs, and thinking about components of the teaching-learning process influence their classroom practices (Kane, Sandretto, & Heath, 2002). What teachers know and believe play a significant role in how they interpret new information and experience, and these interpretations, in turn, guide their instructional practices (Phipps & Borg, 2009). To clarify the overlapping constructs of teacher knowledge, beliefs,

and thinking, Thompson (1992, p. 130) proposed the term "conceptions" to refer to "a more general mental structure, encompassing beliefs, meanings, conceptions, propositions, rules, mental images, preferences, and the like." Thus, conceptions are proposed to be organizing frameworks through which individuals understand, respond to, and interact with a phenomenon (Brown, 2004).

One important component of the teaching-learning process that is guided by teachers' conceptions is classroom assessment. Nitko and Brookhart (2007) define assessment as the process of obtaining information using various methods and interpreting these information to make decisions about students. Assessment serves many important purposes in the learning process. When done correctly, it may be used to certify student learning, improve learning and teaching, and even help students to evaluate their own work (Shepard, 2000). Thus, teachers' conceptions about assessment must be examined because they strongly influence how teachers design, implement, and interpret (the results of) classroom assessments.

Teachers' Conceptions of Assessment

Brown (2004) proposed four general conceptions of assessment based on its perceived purpose: (1) Assessment improves learning and teaching; (2) Assessment makes students accountable for learning; (3) Assessment is used to hold schools and teachers accountable; and (4) Assessment is irrelevant to teaching and learning. The first three conceptions of assessment were identified by Brown (2002) through literature review. The last conception was proposed by Brown (2002) after noticing that many teachers feel that assessment, particularly standardized assessment, can be detrimental to teacher autonomy and professionalism, and can disrupt student learning.

To test this four-dimensional framework of teachers' conceptions of assessment, Brown (2004) developed a self-report attitude inventory that measures teachers' conceptions of assessment using a six-point agreement rating scale. The original instrument consisted of 65 items, but was later trimmed to 50 items after exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The resulting Conceptions of Assessment (COA-III) inventory was first tested among New Zealand primary school teachers where it was found that a four-factor model best represented teachers' conceptions of assessment. The four-factor model has also been confirmed with Australian primary and secondary school teachers (Brown, Lake & Matters, 2011). Attempts to test the conceptions of assessment model with teachers from other countries have

shown that this four-factor structure is not invariant. A study conducted by Barnes, Fives & Dacey (2017) showed that the four-factor model did not adequately represent US K-12 teachers' conceptions of assessment. Instead, exploratory factor analysis showed that US teachers have three dominant conceptions of assessment - assessment as valid for accountability, assessment as a tool for improving teaching and learning, and assessment as irrelevant (Barnes et al., 2017). Similarly, Fletcher, Meyer, Anderson, Johnston, and Rees (2012) tested the conceptions of assessment model with higher education faculty from four New Zealand universities, and found that a two-factor higher-order model was a better fit compared to Brown's (2004) original four-factor model. Their analysis showed that higher education faculty generally hold either a positive conception of assessment that focuses on improvement or a negative conception of assessment that considers assessment to be irrelevant to their practice.

Departing from a quantitative approach to examining conceptions of assessment, Remesal (2011) did a qualitative study with primary and secondary teachers in Barcelona, Spain. The study examined teachers' conceptions about assessment in terms of four aspects: the learning process, the teaching process, accreditation of learning, and accountability of teachers. Remesal (2011) proposed a four-dimensional bipolar model of conceptions of assessment. In this model, teachers' conceptions of assessment fall into two categories: the pedagogical-regulation pole, which focuses on the monitoring of teaching and learning, and the societal-accreditation pole, which focuses on teachers' accountability and certification of achievement. Remesal (2011) noted, however, that these conceptions do not appear to be mutually exclusive, but may occur in different combinations (extreme pedagogical, mixed pedagogical, mixed societal, and extreme societal). This suggested that teachers may hold contradictory beliefs about how assessment affects teaching and learning and that their conceptions of assessment may not be organized as neatly as Brown's four-factor model (2004).

Classroom Assessment in the Philippines

The Department of Education (2015) emphasized that classroom assessment is an important component of curriculum implementation as it allows teachers to track and measure students' progress and to adjust instruction appropriately. The department's policies and guidelines on classroom assessment are articulated in the DepEd Order No. 8, which has been implemented since School Year 2015-2016. In this document, classroom

assessment is classified as either formative or summative. Formative assessment is defined as either "assessment FOR learning so teachers can make adjustments in their instruction" or as "assessment AS learning wherein students reflect on their own progress" (Department of Education, 2015, p. 3). The memo further described formative assessment as "characteristically informal" and "intended to help students identify strengths and weaknesses in order to learn from the assessment experience" (Department of Education, 2015, p. 3). Their description also emphasized that formative assessment is something that the teacher gives to the student at any time during the teaching and learning process in order to track student progress and make informed instructional decisions. Summative assessment, on the other hand, is described as "assessment OF learning, which occurs at the end of a particular unit" (p. 3) to judge whether or not students have met the content and performance standards specified in the K-12 curriculum (Department of Education, 2015). The memo also noted that the results of summative assessments are recorded and reported to the learners, their parents or guardians, and to school administrators.

It seems that the Department of Education views assessment as serving dichotomous functions - formative, which focuses on using assessment to improve learning and teaching, and summative, which certifies student learning. Remesal (2011) noted that a dichotomous perspective of the functions of assessment is limited and does not sufficiently capture the complexity of the assessment process. Moreover, the Department of Education's description of formative assessment as a process that teachers perform upon their students fails to recognize the role and responsibility of the students in the learning process. According to Black and Wiliam (2009), formative assessment must include the learners as owners of their learning and as instructional resources for each other. Thus, Filipino teachers might hold conceptions of assessment that are narrow or limited, considering that the Department of Education has explicitly stated such views in their policy guidelines. This notion, however, needs to be verified as there seems to be a dearth of studies on assessment beliefs and practices of Filipino teachers.

One such study examined the assessment literacy of Filipino pre-service and in-service teachers (Balagtas, Dacanay, Dizon, & Duque, 2010). Using a self-report survey, the study found that Filipino pre-service and in-service teachers are weak in terms of the following competencies: administering, scoring, and interpreting results of externally-produced and teacher-made tests; using assessment results to make different types of academic decisions; developing valid student grading procedures; communicating assessment results to students, parents, and other stakeholders; and recognizing unethical, illegal,

and other inappropriate methods and uses of assessment. Furthermore, the study found that teachers are unprepared to use alternative assessments, as they were not trained to do so.

The Present Study

Given that teachers' beliefs are believed to influence classroom practices and outcomes, this study attempts to examine Filipino teachers' conceptions of assessment while addressing the limited literature on classroom assessment in the Philippine context. This is an initial study that seeks to test the structure of the Conceptions of Assessment Inventory (COA-III; Brown, 2004) among Filipino teachers. Previous studies on conceptions of assessment have shown that the factor structure of the COA-III varies across samples (e. g., Barnes et al., 2017; Brown, 2004; Brown & Remesal, 2012; Fletcher et al., 2012). The present study seeks to determine if Filipino teachers' conceptions of assessment may be represented adequately by the model of Brown (2004). If the model is found to be appropriate for Filipino teachers, then the model can be used to inform professional development programs and assessment decisions in the classroom.

Method

Participants

A total of 391 Filipino teachers (74% females, 26% males) participated in the present study. Eighty-seven percent (n = 342) teach basic education, and 13% (n = 49) teach in the tertiary level. The mean age of the respondents is 32.89 years, with a standard deviation of 10.94 years. On the average, the participants' length of teaching experience is 9.07 years (SD = 9.23 years). Majority (n = 330 or 84%) of the participants have obtained a Bachelor's degree, but only a handful have completed higher degrees (n = 56 or 14% have completed their Master's degrees and n = 5 or 2% hold a Doctorate degree). The participants teach various subjects: Mathematics (n=63 or 16%), Science (n=59 or 15%), English (n=56 or 14%), different subjects in self-contained classes (n=55 or 14%), professional courses (n=42 or 11%), Filipino (n=24 or 6%), Social Studies (n=24 or 6%), Music, Arts, PE and Health (MAPEH) (n=21 or 5%), Christian Living/Religious Education (n=17 or 4%), and Technology

and Livelihood Education/Computer (n=10 or 3%). There were 20 teachers (5%) who did not indicate the subjects they are teaching.

Instrument

Filipino teachers' beliefs about assessment were measured using the Conceptions of Assessment Inventory III (COA-III; Brown, 2004). The COA-III is a self-report attitude inventory that measures teachers' degree of agreement or disagreement with statements related to assessment. The scale consists of 50 items that load on two first-order factors (School Accountability and Student Accountability), and on two second-order factors (Improvement, with four subscales, and Irrelevance, with three subscales). The COA-III subscales, their key premises, and sample items are shown in Table 1. The COA-III uses a 6-point positively-packed agreement rating scale (strongly disagree, mostly disagree, slightly agree, moderately agree, mostly agree, strongly agree), which has been shown to be appropriate when respondents are likely to hold positive attitudes toward the construct being measured (Brown, 2004).

Data Analysis

Means and standard deviations were calculated for the nine subscales of the COA-III. Measures of skewness and kurtosis were also computed to provide more information about how the subscale scores were distributed. Cronbach's alpha values were also calculated for each of the subscales to establish internal consistency reliability.

The factor structure of the COA-III was tested using a combination of exploratory and confirmatory factor analyses. First, confirmatory factor analysis (CFA) using IBM SPSS AMOS was used to test Brown's (2004) original four-factor model with two first-order factors (school accountability and student accountability) and two second-order factors (improvement and irrelevance). The following fit indices and criteria were used to establish model fit: Chi-square index is statistically non-significant; the root-mean-square-error-of-approximation (RMSEA) is .06 or less; standardized root mean square residual (SRMR) is .08 or less; and the comparative fit index (CFI) and the Tucker-Lewis index (TLI) are at least .95 (Hu & Bentler, 1999).

Table 1
COA-III Subscales and Sample Items

First-Order Factor	Second-Order Factor	Key Premise	Number of items
School Accountability		Assessment can be used to account for a teacher's, a school's, or a system's use of society's resources.	6
Student Accountability		Assessment is used to hold students accountable for their learning.	7
Improvement	Describe	Assessment is used to describe the abilities, knowledge, and thinking of students.	6
Improvement	Student Learning	Assessment improves student learning.	7
Improvement	Validity	Assessment information is valid.	5
Improvement	Teaching	Assessment improves teaching.	6
Irrelevance	Bad	Assessment is bad for teaching.	5
Irrelevance	Ignore	Teachers may use assessment, but they ignore it.	5
Irrelevance	Accurate	Assessment is inaccurate.	3

Considering that previous studies (e. g., Barnes et al., 2017; Brown & Remesal, 2012; Fletcher et al., 2012) yielded findings that were not consistent with Brown's (2004) four-factor second-order model of conceptions of assessment, the present study also followed an exploratory approach. The purpose of the exploratory factor analysis (EFA) was to discover the underlying dimensions of Filipino teachers' conceptions of assessment without specifying a priori dimensions. This approach was used by Brown and Remesal (2012) in examining the structure of Spanish teachers' conceptions of assessment.

Exploratory factor analysis was conducted using SPSS. Principal axis factoring (PAF) with promax rotation was used to extract the factors that would best represent Filipino teachers' conceptions of assessment. PAF is recommended when the purpose of the analysis is not simply to reduce data, but to express the relationships among the items in a scale in terms of their underlying latent dimensions (Floyd & Widaman, 1995). Also, Fabrigar, Wegener, MacCallum, & Strahan (1999) noted that PAF may be used even when the assumption of multivariate normality is violated. Four criteria were used for deciding on how many factors to retain: (1) Kaiser-Guttman criterion (i.e., retain factors with eigenvalues greater than 1); (2) the scree test (i.e., number of factors to retain is determined by locating the point in the scree plot at which the slope is zero); (3) factors that have at least three items with factor loadings greater than 0.3; and (4) interpretability of the resulting factor structure (Fabrigar et al., 1999). The obtained factor structure was then rotated using an oblique rotation method (i. e., promax rotation) because teachers' conceptions of assessment are thought to have dimensions that would be related to each other. The underlying dimensions found using exploratory factor analysis were then compared with the four-factor model proposed by Brown (2004).

Results

Descriptive Statistics

The mean, standard deviation, skewness, and kurtosis of each COA-III subscale were calculated to determine the distribution of scores. Cronbach's alpha values of each subscale were also determined. These statistics are all presented in Table 2.

Table 2
Descriptive Statistics for the Conceptions of Assessment Inventory (COA-III) (n =391)

COA-III Subscales	*M*	*SD*	Skewness	Kurtosis	Cronbach's α
School Accountability	4.28	0.98	-0.53	0.01	0.87
Student Accountability	4.19	0.88	-0.69	0.62	0.77
Improvement	4.58	0.88	-1.26	1.91	0.95
Describe	4.83	1.08	-0.48	5.79	0.86
Student Learning	4.79	0.97	-1.35	1.99	0.89
Validity	4.12	0.90	-0.65	0.30	0.71
Teaching	4.61	0.97	-0.70	1.85	0.84
Irrelevance	3.83	0.82	-0.56	0.66	0.81
Bad	4.13	1.03	-0.31	0.00	0.64
Ignore	4.45	1.20	-0.89	0.34	0.83
Accurate	2.93	0.93	0.25	0.55	0.37

The rating scale used for the COA-III is a six-point positively-packed agreement rating scale, with mean scores ranging from 2.93 to 4.83. The standard deviations show minimum dispersion of scores from the mean of each subscale. The skewness values range from -1.26 to 0.25, but most of the values are negative, indicating that the scores for the subscales are concentrated on the higher values. Subscales whose skewness values are between -0.5 to 0.5 indicate that the distribution of scores is slightly skewed, while those whose skewness values are smaller than -1.0 or larger than 1.0 indicate a distribution that is moderately skewed. This is to be expected since the scale that was used is positively packed (i.e., two negative responses and four positive responses). Positively packed scales have been shown to be effective in generating a variety of responses when participants are inclined to respond positively to items (Brown, 2004). The positive kurtosis values indicate that the distribution is leptokurtic (i.e., more peaked than the normal distribution and has fatter tails).

Confirmatory Factor Analysis

Brown (2004) proposed that the structure of teachers' conceptions of assessment may be represented by a four-factor model with two first-order

factors (School Accountability and Student Accountability) and two second-order factors (Improvement and Irrelevance). This four-factor second-order model was tested among a heterogenous group of Filipino teachers.

All of the items in Brown's (2004) COA-III loaded significantly on their hypothesized factor ($p < .05$), except for two items under Irrelevance-Accurate (i.e., *Assessment is an imprecise process* and *Assessment results should be treated cautiously because of measurement error*) which had non-significant factor loadings. These two items had standardized factor loadings below .4, which indicates that these items seem to be unrelated to the underlying dimension they are supposed to measured. Table 3 shows the standardized factor loadings and standard errors of each item in the COA-III.

Table 3
COA-III Statements, Factor Loadings, and their Standard Errors

Factors and Statements	Unstandardized Factor Loading	Standardized Factor Loading	Standard Error
School Accountability			
Assessment measures the worth or quality of schools	1.000	.654	
Assessment is an accurate indicator of a school's quality	1.155	.764	.089
Assessment shows the value schools add to student learning	.924	.775	.070
Assessment is a good way to evaluate a school	1.126	.803	.083
Assessment influences the way teachers think	1.085	.806	.080
Assessment keeps schools honest and up-to-scratch	.664	.514	.072
Student Accountability			
Assessment selects students for future education or employment opportunities	1.000	.487	
Assessment is comparing student work against set criteria	.926	.467	.128
Assessment determines if students meet qualifications standards	1.410	.838	.142

Factors and Statements	Unstandardized Factor Loading	Standardized Factor Loading	Standard Error
Assessment is assigning a grade or level to student work	1.108	.606	.130
Assessment places students into categories	1.172	.621	.136
Assessment is checking off progress against achievement objectives	1.096	.620	.127
Assessment is completing checklists	.538	.258	.118
Improvement-Describe			
Assessment is a way to determine how much students have learned from teaching	1.000	.790	
Answers to assessment show what goes on in the minds of students	.902	.716	.058
Assessment measures students' higher order thinking skills	1.048	.809	.057
Assessment establishes what students have learned	1.015	.845	.052
Assessment identifies student strengths and weaknesses	1.100	.879	.054
Assessment identifies how students think	.810	.273	.150
Improvement-Student Learning			
Assessment feeds back to students their learning needs	1.000	.769	
Assessment helps students improve their learning	.967	.809	.055
Assessment is appropriate and beneficial for children	1.001	.815	.057
Assessment provides feedback to students about their performance	1.001	.862	.053
Assessment is an engaging and enjoyable experience for children	.636	.508	.062
Assessment makes students do their best	.948	.767	.058
Assessment is a positive force for improving social climate in a class	.758	.589	.063
Improvement-Validity	1.000	.695	

Factors and Statements	Unstandardized Factor Loading	Standardized Factor Loading	Standard Error
Assessment results are trustworthy			
Assessment results predict future student performance	1.020	.621	.095
Assessment results are consistent	.539	.327	.093
Assessment is objective	.887	.575	.089
Assessment results can be depended on	.928	.610	.088
Improvement-Teaching			
Assessment influences the way teachers think	1.000	.608	
Assessment is integrated with teaching practice	1.287	.842	.099
Assessment changes the way teachers teach	1.172	.682	.104
Assessment information modifies ongoing teaching of students	1.134	.712	.098
Assessment allows different students to get different instruction	1.238	.507	.140
Assessment information is collected and used during teaching	1.257	.695	.110
Irrelevance-Bad			
Teachers pay attention to assessment only when stakes are high	1.000	.496	
Assessment interferes with teaching	.646	.278	.139
Teachers are over-assessing	1.197	.671	.140
Assessment is unfair to students	1.305	.721	.147
Assessment forces teachers to teach in a way against their beliefs	.709	.344	.128
Irrelevance-Ignore			
Teachers ignore assessment information even if they collect it	1.000	.729	
Assessment has little impact on teaching	.893	.616	.078
Teachers conduct assessments but make little use of the results	.842	.635	.071
Assessment is value-less	.977	.752	.070

Factors and Statements	Unstandardized Factor Loading	Standardized Factor Loading	Standard Error
Assessment results are filed and ignored	1.076	.797	.073
Irrelevance-Accurate			
Teachers should take into account the error and imprecision in all assessment	1.000	1.618	
Assessment is an imprecise process	.021	.028	.079
Assessment results should be treated cautiously because of measurement error	.152	.246	.518

It should also be noted that four first-order factors (Describe, Student Learning, Validity, and Teaching) had significant factor loadings (standardized β's > .85) on their second-order factor (i.e., Improvement). However, of the three first-order factors (Bad, Ignore, and Accurate), only two factors (Bad and Ignore) loaded significantly (standardized β's > .90) on their second-order factor (i.e., Irrelevance), indicating that the first-order factor Irrelevance-Accurate may be unrelated to its hypothesized second-order factor.

Correlations among the second-order factors are shown in Table 4. Large (*r's* > .50) effect sizes were observed among school accountability, student accountability, and improvement. Large effect sizes (*r's* > .50) indicate that the proportion of variance shared by school accountability, student accountability, and improvement is at least 25%. Take for instance, school accountability and student accountability have an observed $r = .73$, which indicates that 53.29% of the variation in school accountability may be explained by the variation in student accountability. This simply means that differences in Filipino teachers' beliefs pertaining to assessment as a means for ensuring school accountability may be accounted for by the differences in their beliefs regarding assessment as a means to promote student accountability. This is not to say that the two factors are similar. Rather, these factors vary together, and in the same direction, suggesting that when teachers' hold positive beliefs about assessment as necessary for school accountability. They also tend to hold positive beliefs about assessment as important for ensuring student accountability. The observed positive relationships among these factors also provide evidence for convergent validity.

Moreover, the observed indirect relationship between school accountability and irrelevance (r = -.18), and between student accountability and irrelevance (r = -.23) provide support for discriminant validity. This means that if teachers think of assessment as helpful for holding schools and students accountable, then they most likely will not consider assessment as irrelevant. However, the observed effect sizes (r's < |.10|) were small, which means that the proportion of variance in school accountability and student accountability that may be explained by the variation in irrelevance is less than 5%. It should also be noted that the observed effect size (r's = .01) between improvement and irrelevance is not significant, which means that these factors are unrelated.

Table 4
COA-III Second-Order Factor Correlations

Factor	(1)	(2)	(3)	(4)
(1) School Accountability	-			
(2) Student Accountability	.73*	-		
(3) Improvement	.76*	.75*	-	
(4) Irrelevance	-.18*	-.23*	.01	-

* Correlation is significant at the 0.05 level (2-tailed)

The four-factor second-order model of the COA-III showed poor fit: χ^2 = 3,617.089, df = 1162, p-value = .000; χ^2 /df = 3.113; RMSEA = 0.074; SRMR = 0.1179; CFI = 0.784; TLI = 0.772. A generally accepted procedure in CFA is to respecify the model by removing items that have low factor loadings or by allowing error terms to be correlated. This modified model is then tested (see Byrne, 2010 for an illustration). This procedure is more appropriate, however, if the second (i.e., respecified) model were to be tested on a sample that is different from the one that was used to derive the first model (Sharma, 1996). A common practice is to use a relatively large sample, split it into two subsamples – a derivation sample and a cross-validation sample, and use the cross-validation sample to test the modified model (Floyd & Widaman, 1996). In this study, however, the obtained sample size (n = 391), was adequate only for deriving a model, but not for a cross-validation. Thus, we opted not to respecify and test a second model at this time for lack of an adequate cross-validation sample.

The obtained fit indices indicate that Filipino teachers' responses on the COA-III did not sufficiently support the four-factor model of conceptions of

assesment proposed by Brown (2004). This also suggests that the Filipino teachers may hold different conceptions of assessment that are not captured by Brown's model. The results of the CFA also lend support to our initial proposal of doing an EFA in order to discover the underlying structure of Filipino teachers' conceptions of assessment.

Exploratory Factor Analysis

Previous research has shown that the factor structure of the COA-III (Brown, 2004) varies across samples from different countries (Barnes et al., 2017; Brown & Remesal, 2012; Fletcher et al., 2012). There is empirical evidence, therefore, to support the notion that the structure of Filipino teachers' conceptions of assessment may be different from that of New Zealand teachers who composed the sample for the original inventory. Thus, the current study explored the underlying dimensions of Filipino teachers' conceptions of assessment without specifying a priori the structure of these dimensions.

To determine whether the data are adequate for an EFA, we first examined the correlations among the 50 items of the COA-III. We found moderate correlations, indicating that the items may be grouped into homogeneous factors that measure the similar underlying dimensions. Second, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was found to be .948. According to Sharma (1996), a KMO value of .90 or higher means that the items are homogeneous enough and, therefore, appropriate for factoring. Third, Bartlett's test of sphericity resulted to a p-value less than $a = .05$, which indicates that there is redundancy between items that can be summarized into factors, and therefore factor analysis is appropriate (Sharma, 1996).

Exploratory factor analysis using principal axis factoring with oblique (promax) rotation revealed four factors with eigenvalues greater than one. The scree plot, however, showed that there are three factors above the inflection point, which suggests that only three factors ought to be retained based on this heuristic. Since the two criteria (i.e., Kaiser-Guttman and scree plot) resulted to inconsistent results, we examined the items that loaded significantly on the extracted factors. Closer inspection of the factor loadings showed that for the fourth factor, there were only two items that had factor loadings greater than .3, which suggests that the fourth factor may no longer be interpretable. Thus, we decided that only three factors would be retained. These three factors explained 45.45% of the variance. Table 5 shows the eigenvalues, the percentage of values accounted for by each factor before and after rotation, and the number of items in each factor that had factor loadings of 0.3 or higher.

Table 5
Eigenvalues, Variance Explained and Items Included for the Three-Factor Model

Factor	Eigenvalues	Percentage of Variance Explained		No. of Items
		Unrotated	Rotated	
1	16.878	33.756	33.023	17
2	5.450	10.900	9.963	8
3	1.678	3.356	2.464	6

It should be noted that the items did not load on the original factors conceptualized by Brown (2004). Also, only 31 items were retained from the original scale because these were the items that had factor loadings greater than 0.3. We then examined the resulting factor structure to see if it was interpretable and theoretically sensible, as suggested by Fabrigar et al. (1999). Table 6 shows the COA-III items that were retained for a three-factor solution, their factor loadings, and the Cronbach's alpha coefficients of the new subscales.

There were 17 items that loaded on the first factor, which explained 33.023% of the variance after rotation. Fifteen of these items measured the Improvement function of assessment in the original COA-III inventory. Two items focused on Student Accountability (*Assessment determines if students meet qualifications standards* and *Assessment is checking off progress against achievement objectives*). These items, however, seem to be still related to the notion that assessment serves an improvement function, because it also implies that they have improved when students meet qualification standards or show progress against achievement objectives. Thus, we named Factor 1 as the *Improvement* factor, similar to how Brown's (2004) and Barnes et al.'s (2017) interpretation. Since there were 17 items that loaded on this factor, we also checked if there would be a second-order factor by running another EFA using the 17 items. Only one factor was extracted and it explained 59.043% of the variance, suggesting unidimensionality. Thus, we retained all 17 items under the Improvement factor.

Table 6
Factor Loadings of the Oblique, Three-Factor, 31-item Solution with Internal Consistency Coefficients

Factor / Item	Factor			Cronbach's
	1	2	3	α
Factor 1 – Improvement				**0.951**
21. Assessment provides feedback to students about their performance	0.966			
22. Assessment identifies student strengths and weaknesses	0.935			
45. Assessment helps students improve their learning	0.812			
12. Assessment feeds back to students their learning needs	0.798			
26. Assessment establishes what students have learned	0.774			
48. Assessment is a way to determine how much students have learned from teaching	0.77			
35. Assessment is integrated with teaching practice	0.758			
20. Assessment determines if students meet qualifications standards	0.733			
41. Assessment influences the way teachers think	0.712			
6. Assessment information is collected and used during teaching	0.664			
32. Assessment measures students' higher order thinking skills	0.635			
38. Assessment is appropriate and beneficial for children	0.593			
9. Assessment makes students do their best	0.582			
28. Assessment is checking off progress against achievement objectives	0.566			
18. Assessment information modifies ongoing teaching of students	0.534			
33. Assessment changes the way teachers teach	0.522			
7. Assessment allows different students to get different instruction	0.41			
Factor 2 – Irrelevance				**.862**
43. Teachers conduct assessments but make little use of the results		0.756		
23. Teachers ignore assessment information even if they collect it		0.721		
50. Assessment results are filed and ignored		0.696		

8. Teachers are over-assessing	0.682		
1. Teachers pay attention to assessment only when stakes are high	0.641		
46. Assessment is value-less	0.595		
40. Assessment has little impact on teaching	0.525		
11. Assessment is unfair to students	0.502		
Factor 3 – School Accountability			**0.853**
42. Assessment provides information on how well schools are doing		0.835	
37. Assessment is a good way to evaluate a school		0.810	
30. Assessment is an accurate indicator of a school's quality		0.760	
17. Assessment measures the worth or quality of schools		0.448	
27. Assessment places students into categories		0.378	
24. Assessment results predict future student performance		0.333	

The second factor had eight items and explained 9.963% of the variance after rotation. Items in this factor pertain to the notion that assessment is irrelevant to the teaching and learning process; thus, we named the factor *Irrelevance*, similar to how Brown (2004) and Barnes et al. (2017) interpreted the same results. Finally, the third factor contained five items that describe assessment as necessary for holding schools accountable and one item that describe assessment as a means for predicting future student performance, which is related to the notion that assessment information is valid. However, this item may still be considered as related to the school accountability function of assessment. It also indicates that the school has been doing well in making good use of its resources when students perform well into the future. Thus, we labeled the third factor as *School Accountability*. This factor accounted for 2.464% of the variance after rotation.

To verify this three-factor model, we ran a second EFA with a fixed three-factor solution. The second EFA yielded a three-factor model that explained 44.90% of the variance and had a factor structure similar to the initial model. The first factor that was extracted in the second EFA included items that were related to the *Improvement* function of assessment. Sixteen of the 17 items that loaded on the first factor in the initial EFA also loaded highly on the first factor in the second EFA (i.e., fixed three-factor solution). The second factor extracted in the second EFA had items that were related to the *School Accountability* function of assssment. In the first EFA, the School Accountability

factor was the third factor extracted and had six items. Of these six items, five also loaded highly on the Accountability factor of the second EFA. The third factor extracted in the second EFA included items related to the *Irrelevance* conception of assessment. These same items also loaded on the *Irrelevance* factor in the first EFA. The results of the second EFA with a fixed three-factor solution provide tentative support for our decision to retain three factors in the initial solution. We recognize, however, that since we used the same data set to run a second EFA, our findings need to be verified using an independent sample. Nonetheless, the second EFA suggests that a three-factor model of conceptions of assessment may be more functional for Filipino teachers than Brown's (2004) original four-factor model.

To provide evidence of the functionality of the three-factor solution, we also examined factor intercorrelations. As expected, the Improvement and School Accountability factors are significantly correlated and the effect size is large ($r > .5$). Also, there is no significant relationship between the Irrelevance and School Accountability factors. These results are consistent with the findings of Brown (2004).

Table 7
Correlations among the Three Factors of the COA-III

Factor	(1)	(2)	(3)
(1) Improvement	-		
(2) Irrelevance	.322**	-	
(3) School Accountability	.679**	-0.028	-

**Correlations are significant at the .001 level. (2-tailed)

An interesting result is the positive and significant relationship between Improvement and Irrelevance ($r = .322$). This seems to be counter intuitive because when assessment is seen to improve learning and teaching, then it should be considered as relevant, as explained by Brown (2004). However, it is also likely that, although teachers conceive of assessment as something that is helpful in improving learning and teaching, they may still think of some aspects of assessment as irrelevant, particularly when assessment results are not put to good use. An illustration of this mixed conceptions of assessment may be seen in a qualitative study of Spanish teachers' conceptions of assessment conducted by Remesal (2011).

Remesal (2011) found that teachers's views about assessment belong to two types: pedagogical (focusing on the monitoring of teaching and learning) and societal (focusing on teachers' accountability and the certification of achievement). However, Remesal (2011) pointed out that these two sets of beliefs are not mutually exclusive, and that teachers may hold a combination of these beliefs – a purely pedagogical conception, a purely societal conception, a mixed pedagogical conception, and a mixed societal conception. Thus, the observed positive and significant relationship between the Irrelevance and Improvement factors provide support for Remesal's (2011) contention that a strictly dichotomous interpretation of the functions of assessment (positive vs. negative or formative vs. summative) may not be functional or realistic. Teachers may indeed hold contradictory beliefs about the importance and relevance of assessment in teaching and learning. This means that while teachers believe that assessment is important in the teaching and learning process, current practices in and expectations from the educational system may lead them to also think of assessments as irrelevant. For instance, teachers in basic education may use assessments to facilitate student learning and enhance classroom instruction. In this case, assessment serves an improvement function. However, these same teachers also need to prepare their students to take high-stakes tests, such as the National Achievement Test prescribed by the Department of Education. Thus, classroom assessments may no longer serve a pedagogical function when they are meant only to prepare students to get high scores in standardized tests, which are often used to rank schools, particularly among public schools. Private schools, on the other hand, may require teachers to develop a variety of assessments in compliance with the Department of Education (2015) policy guidelines on classroom assessment. The policy guidelines prescribe specific types of assessments (i.e., written works, performance tasks, and quarterly assessment) that teachers must include in rating student performance, whether or not these assessments are aligned with the learning outcomes or with teaching and learning activities. When teachers develop assessment tasks out of compliance, they might consider assessment to be irrelevant or unimportant to the teaching and learning process.

The situation in higher education may not be any different from that of the basic education sector. Since higher education is intended to prepare students for a particular profession, there is often a great deal of pressure for university professors to cover the necessary content that students need to know. Thus, assessments in the university level may be important only as far as ensuring that students have gained the necessary knowledge and skills so that they can be ready to take on roles related to their chosen profession. The

pedagogical purpose of assessment (i.e., to help students improve their learning, to facilitate student engagement; to improve classroom instruction) may be overlooked when the focus of assessment is to certify student achievement of desired knowledge and skills.

Discussion

The current study tested the factor structure of the COA-III inventory (Brown, 2004), an instrument designed to measure teachers' conceptions about the functions of asssessment. Using a combination of CFA and EFA, we found support for findings of previous studies about the lack of invariance of the scale across samples from different countries (see Brown & Remesal, 2012; Fletcher et al., 2012; and Barnes et al., 2017). We also found that Filipino teachers may hold contradictory beliefs about the functions of assessment.

These findings have implications for the training of both preservice and in-service teachers in assessment. Teachers' beliefs strongly influence their actual classroom practice, therefore, it is important for teacher education institutions to incorporate in their assessment courses some opportunities to examine their beliefs. Teacher educators need to address students' misconceptions or faulty understandings of assessment. In-service training of teachers must also include sessions on assessment that would focus not just on the strategies for classroom assessment, but also on examining teachers' beliefs about the purpose and functions of assessment.

Considering that the structure of Filipino teachers' conceptions of assessment seems to be misrepresented by the current model of Brown (2004), it would benefit teacher educators, school administrators, policy makers, and the teachers themselves if a more indigenous model would be developed. Remesal (2011) pointed out that the assessment beliefs of teachers may be affected by the structure of the educational system. This means that in developing an indigenous model of Filipino teachers' conceptions of assessment, their context, experiences, and challenges within the educational system must also be considered. Future research, therefore, may be undertaken using either a qualitative or a mixed-methods approach to determine a model that would more closely represent their conceptions of assessment. Eventually, such model may be used to inform teacher training programs and to examine assessment practices in the classroom.

Finally, our study is an attempt to address the dearth of research on classroom practices of Filipino teachers. It is our hope that this and future

research will not only help Filipino teachers in improving their practice, but will also give them voice in the literature on educational assessment.

References

Balagtas, M. U., Dacanay, A. G., Dizon, M. A., & Duque, R. E. (2010). Literacy level on educational assessment of students in a premiere teacher education institution: Basis for a capability building program. *The Assessment Handbook*, *4*, 1-19.

Barnes, N., Fives, H., & Dacey, C. M. (2017). US teachers' conceptions of the purposes of assessment. *Teaching and Teacher Education*, *65*, 107-116.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*, 5-31.

Brown, G. T. L. (2002). *Teachers' conceptions of assessment* (Auckland, NZ, University of Auckland).

Brown, G. T. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*, 301-318.

Brown, G. T. (2006). Teachers' conceptions of assessment: Validation of an abridged version. *Psychological reports*, *99*(1), 166-170.

Brown, G. T., & Remesal, A. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. *The Spanish Journal of Psychology*, *15*, 75-89.

Brown, G. T., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, *27*, 210-220.

Department of Education. (2015). *Policy guidelines on classroom assessment for the K-12 Basic Education Program*. Manila, Philippines.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Fletcher, R. B., Meyer, L. H., Anderson, H., Johnston, P., & Rees, M. (2012). Faculty and students conceptions of assessment in higher education. *Higher Education*, *64*, 119-133.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286-299.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*, 1-55.

Kane, R., Sandretto, S., & Heath, C. (2002). Telling half the story: A critical review of research on the teaching beliefs and practices of university academics. *Review of Educational Research*, *72*(2), 177-228.

Nitko, A. J., & Brookhart, S. M. (2007). Educational assessment of student. Englewood Cliffs. NJ: Merrill Prentice Hall.

Phipps, S., & Borg, S. (2009). Exploring tensions between teachers' grammar teaching beliefs and practices. *System*, *37*, 380-390.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, *27*, 472-482.

Sharma, S. (1996). *Applied multivariate techniques*. New York: John Wiley & Sons.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*, 4-14.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 127-146). New York, NY, England: Macmillan Publishing Co, Inc.

Educational Measurement and Evaluation Review (2018), Vol. 9 Issue 1

Development and Validation of *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition) of the Filipino Reading Achievement Test

Ryan Francis O. Cayubit
Lyka Ilonah D.C. Chua
Emerald Ann S. David
Therese Monique D.G. Gutierrez
Shiara Marriz T. Marquez
Niko A. Mendoza
Emille Joyce P. Palogan
Reniel B. Tiu
University of Santo Tomas

Abstract

The objective of this cross-sectional exploratory study is to develop and standardize two subtest for the proposed Filipino Reading Achievement Test. Subtests measuring *Pagbabaybay* or Spelling in Filipino and *Pagkilala sa Binasa* or Word Recognition was constructed. The focus of the study is assessing the validity and reliability of these instruments in order to justify its use in assessing the reading abilities of Filipino children. Assessing reading ability is essential because reading has been considered as one of the most important skill an individual needs to develop in order to succeed both in school and in the world of work. The study followed the standard scale development procedures of item analysis, reliability and validity testing. Implications of the findings are discussed.

Keywords: reading, spelling, word recognition, achievement test, Filipino children

Introduction

Learning is a never-ending process that is spurred by man's desire to acquire more knowledge and experience. There are several ways to learn and

one of this is through reading. Reading is a linguistic skill that gives meaningful interpretation to printed or written verbal symbols (Tinker & McCullough, 1975). The goal of which is to transfer ideas from written text to the human mind by integrating all materials in order to arrive at something meaningful (Tinker & McCullough, 1975). This unique human skill is acquired after one has gained substantial aptitude in oral language (Muter, Hulme, Snowling, & Stevenson, 2004) and is believed to be enhanced by formal training and education.

One characteristic of reading that sets it apart from other literacy skills is its automaticity (Schwatrz, 2003). Reading automaticity is a fast, accurate and effortless word identification process where an individual reads quickly and automatically while simultaneously keeping a flow of thoughts and ideas that will enable him to generate inferences and establish connections within the text that was read (Hook & Jones, 2002; Warrington, 2006). Although seemingly simple at the surface, a deeper understanding of reading would lead one to realize that it actually requires an instantaneous understanding of the text presented without having to pause for conscious efforts to decode the letters and put them together to ascribe meaning to the words, sentences and ultimately, the entire text presented. This implies that the reader has the ability to use the different components of reading like vocabulary knowledge (understanding of the meanings of individual words), word recognition (the immediate identification of common words), and spelling (the ability to relate sounds to symbols in both familiar and unfamiliar words), among others (Muter et al., 2004). These components make up what is known as the reading achievement of an individual.

The theoretical underpinning of reading appears vast and it seems that reading can be best understood if it is not constricted by just one theory. For instance, Piaget's theory of intellectual development, when used side by side with the interaction model of reading leads to a better understanding of reading performance at specific developmental stages (Graves, Watts-Taffe & Graves, 1999; Karlin, 1973). Piaget's theory of intellectual development posits that from the age of 7 to 11, a child has already acquired the skill to logically organize cognitive activities as characterized by the processes important to this stage like composition, associativity, identity, reversibility, and seriation. Composition allows the child to view parts and represent them as contained in the whole, while associativity gives the child the ability to arrange and rearrange elements of what is presented to him in various ways. Identity, on the other hand, lets the child maintain a perception of the original so that he can return it to that

condition when changes take place and reversibility enables the child to coordinate or compare the alterations made among the elements of a group. Finally, seriation allows the child to sequence elements depending on some criterion. All of these falls under what Piaget calls the concrete operational stage (Boyd & Bee, 2015; Karlin, 1973; Santrock, 2017) and are necessary for reading to take place. Meanwhile, the interaction model of reading introduced by Rumelhart in 1977 explains that not only is reading influenced by the child's ability to manipulate what is presented to him but is also affected by the text itself (Graves et al., 1999). This means that an interaction seems to exist between reader characteristics (as presented in Piaget's developmental theory) and properties of the text being read (supplemented by Rumelhart's interactive model). This theoretical interaction is what produces meaning ascribed to the material read by an individual.

Despite the existence of theories and the importance of reading as a contsurct, measuring reading achievement of Filipino children, particularly if cultural characteristics is taken into account, remains a challenging task. It is challenging because most available scales may not be appropriate for Filipino children. This disparity can be traced to discrepancies in the quality of education and many any other socio-cultural factors, like the lack of fluency in the English language, which may affect the results of these scales resulting to inaccurate assessment. It is in this light that the researchers deem it necessary to develop and standardize an assessment tool that would measure Filipino reading achievement.

Nature of Reading

The importance of reading is well articulated in the literature. Reading is important because it provides the learners with access to a great quantity of future experience of the language and it presents a window onto the normal means of continuing one's personal education (Stevens, 1977). It is also through reading that learners would be able to develop a sufficient language base that enables them to produce the spoken or written messages, which they are eager to communicate to others (Rajabi, 2009). In addition, Roe, Smith and Burns (2005) stated that some educators regard reading as a set of interrelated sub-skills that children must master and integrate.

According to bottom-up theorists, reading is a linear process by which readers start with the written text and decode the text word by word, linking the words into phrases and then sentences (Gove as cited in Roe et al., 2005; Gray

& Rogers as cited in De Dabat, 2006; Gunning, 2003; Phakiti, 2006). The bottom-up theory focuses on how readers extract information from the printed page, claiming that readers deal with letters and words in a relatively complete and systematic fashion (Treiman, 2001). A complex task like reading is broken down into their component skills (Gunning, 2003) such as word recognition, spelling, morpho-phonemic processing and morpho-syntactic parsing (Phakiti, 2006). In this view of reading, readers assume a passive role and more focus is given on the printed symbols or text. Reading processes are considered to be completely under the control of the text and had little to do with the information possessed by a reader or the context of discourse (Gao, 2006).

Another theory sees reading as a top-down process and emphasizes the role of the reader (Gao, 2006) and his background knowledge (Bai, 2007; De Dabat, 2006) in the process of extracting meaning. According to this view, reading is not just obtaining meaning from the text but a process wherein the reader connects his background knowledge in deriving at and predicting the meaning of the text. Reading is not a passive mechanical activity but purposeful and rational, dependent on the prior knowledge and expectations of the reader. It is a matter of making sense of written language rather than decoding print to sound (De Dabat, 2006). Readers approach the text with existing knowledge, and work down to read the entire text (Gao, 2006). The reader proves his active role in the reading process by bringing to the interaction his/her available knowledge of the subject, knowledge of and expectations about how language works, motivation, interest and attitudes towards the content of the text. Through the use of cues, the reader forms hypotheses about which words he will encounter and take in only just enough visual information to test his hypotheses.

More recent theories on reading view the process as a combination of the bottom-up and top-down approaches. This interactive model of reading states that reading happens because the reader extracts meaning from the text but that meaning is not only because of the text alone but is attributed to the interaction between the text and the characteristics and knowledge of the reader reading (Erten & Karakas, 2007; Gao, 2006).

Word Recognition and Spelling

Measuring reading achievement involves summing up an individual's performance in the different facets of reading. This is often done in the educational setting where results are used to judge the ability of the student and

as basis for interventions. Two common facets of reading is word recognition and spelling.

Learning to read requires the operation of recognition. Word recognition allows the reader who recognizes a word to match the full print array of that word with an orthographic-phonological representation of that word which was previously stored in memory (Erickson et al., 2008). Word recognition has several aspects. One is the decoding ability of the reader, which involves the use of knowledge in phonics (Gillet, Temple & Crawford, 2008). In decoding, the reader must apply his general knowledge in determining how the letters and clusters of printed words encode sounds (Erickson et al., 2008). Successful word recognition attaches a pronunciation or equivalent phonological representation stored in memory to an orthographic memory of the whole printed word. When word recognition happens very quickly, with no effort exerted, it is called sight word reading or "automatic" word recognition (Ehri, 2005; Erickson et al., 2008). According to Castles and Coltheart (as cited in Muter et al., 2004), a close relationship exists between the development of word recognition skills and phonological awareness. Phonological awareness, also known as phonological sensitivity, includes the ability to recognize, identify, or manipulate any phonological unit within a word (Ziegler & Goswami, 2005). Numerous studies have shown that possessing good phonological awareness skills characterize good readers, whereas possessing poor phonological awareness skills is evident in poor readers (Brady, 1991; Goswami & Bryant, 1990; Scarborough 2001; Wagner & Torgesen, 1987).

Spelling, on the other hand, is a related operation to word recognition that utilizes the process of production. The ability of the reader to read words usually surpasses his ability to spell the words correctly. Similar to word recognition, phonemic awareness or phonemic segmentation and orthographic representation are also key components of spelling (Gillet et al., 2008). It is an essential and complex skill that revolves around the use of one's visual memory, phoneme-grapheme awareness and orthographic and morphophonemic knowledge (Abler & Walshe, 2004; van Hell, Bosman, & Bartelings, 2003). There are two measures of spelling: spelling dictation and spelling recognition. Spelling dictation involves the ability to produce the graphemes corresponding to pronounced words. This type of spelling appears to be associated with good knowledge of sound–symbol correspondences and phonological analysis skills. In contrast, spelling recognition requires children to visually analyze alternative spelling patterns and choose the real word pattern, a skill that might rely more heavily on orthographic or letter pattern knowledge (Zieger & Goswami, 2005).

Both word recognition and spelling tap into orthographic and phonological representations, but possibly with dissimilar demands on the precision of those representations. However, successful word recognition often requires only partial orthographic and phonological representations, whereas spelling requires that complete representations be accessible in order to successfully spell a word (Friend, De Fries, Wadsworth, & Olson, 2007). In describing the relationship between word recognition and spelling, Friend et al. (2007) found that the pattern of phonological errors and orthographic accuracy in spelling was similar to the pattern of errors in word recognition. This would mean that those judged as poor spellers are often poor readers and good spellers are often good readers.

Measures of Reading Achievement

There appears to be a limited number of researches devoted to the development of scales that could be used to assess reading achievement. One such test is the Test of Word Reading Efficiency developed by Torgeson, Wagner and Rashotte in 1999 (Hayward, Stewart, Phillips, Norris, & Lovell, 2008). The test deals mainly on two important aspects of "word level" reading ability, that of sight word reading and phonemic decoding. The rationale of the test is to measure the student's word level reading efficiency. Gentry (2007) developed another instrument designed to measure spelling. It is a brief developmental spelling test of 5 to 10 chosen words designed to measure the strategies used by children in spelling. The strategies are scribbling, random letter, letters for beginning sounds, a letter for each sound and spelling in chunks of phonetic patters.

The Present Study

Recognizing the importance of reading, this study is an extension of a previous work that developed a vocabulary and reading comprehension subtests for the Filipino Reading Achievement Test (Cayubit, 2012). To complete the battery, the present study revolves around the work and procedures used in developing and standardizing the spelling and word recognition subtests. The specific research questions are as follows:

1. What is the composition of the preliminary and polished forms of spelling and word recognition subtests?

2. How valid are the preliminary forms of spelling and word recognition subtests based on exploratory factor analysis?
3. How valid are the polished forms of spelling and word recognition subtests based on the confirmatory factor analysis?
4. How reliable are the preliminary and polished forms of the spelling and word recognition subtests?

Method

Research Design

This research project is cross-sectional exploratory in nature. A cross-sectional exploratory research involves data gathering from respondents in a single point in time with the purpose of developing and instrument that would measure a phenomenon and at the same time explaining the nature of the said phenomenon (Johnson, 2001). The current project aimed to develop and validate an instrument that would measure specific domains that would form part of a pupil's reading achievement. The said domains are *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition). In addition, this is a descriptive normative research endeavor. As explained by Ariola (2006), the descriptive normative method is a type of descriptive research that describes the status of events and people as they exist through the use of standardized instruments with established norms. Applied to the study at hand, it made use of the newly developed instrument on the research respondents to test its usefulness and likewise describe their current condition as regards their achievement in both *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition).

Participants

A total of 1,290 grades 3 and 4 pupils participated in the study. 160 participated in the development of the preliminary form and 1,130 participated in the validation of the polished form. The participants who responded to the preliminary form were from schools in Antipolo, Parañaque and Laguna, while those who responded to the polished form were selected from various schools within Metro Manila. All respondents were selected using convenient sampling.

Measures

To complete the study, four instruments (preliminary and polished forms of *Pagbabaybay* and *Pagkilala sa Salita*) were developed in consultation with teachers who are experts in Filipino and Reading. Items of both tests were written in Filipino and were designed to be suitable for Grades 3 and 4 pupils. The items in the preliminary form were based on the results of the content validation performed by Filipino and Reading teachers while the items of the polished form were based on the analysis of the data from the preliminary form. More specifically, an item will be included in the polished form provided it met the following conditions: (a) factor loading of at least 0.30 in the exploratory factor analysis; (b) index of item discrimination of good or very good; and (c) index of item difficulty in the optimum level. A detailed description of the scales is found in Table 1.

Table 1
Description of test instruments

Type of Test	No. of Items	Description
Preliminary Form Pagbabaybay	59	*Ang pagsusulit na ito ay sumusukat sa kakayahan ng sumasagot na magbaybay ng pantig ng salitang Filipino Ang mga salitang ito ay karaniwan sa mga mag aaral o bata sa ikaw tatlo at apat na antas.*
Polished Form Pababaybay	27	
Preliminary Form Pagkilala sa Salita	70	*Ang pagsusulit na ito ay sumusukat sa kakayahan ng sumasagot na kilalanin ng tama at walang hirap ang mga salita.*
Polished Form Pagkilala sa Salita	28	

Procedure

Data gathering started with documentary analysis using textbooks on Reading and Filipino, the purpose of which is to generate initial items that would make up the preliminary form. Available reading tests were also reviewed to see how test items were constructed and how they were administered. This was followed by item construction using the common sight words in Filipino as basis. In addition, Piaget's theory of intellectual development and Rumelhart's Interaction Model of Reading also guided the item construction. The initial sets of items were then reviewed by selected experts for content validation. The preliminary forms of *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition) were then administered, scored and collated to determine their initial psychometric properties. The initial psychometric properties served as a guide in revising the preliminary form, items were then retained, revised and discarded based on the criteria mentioned in the previous section. The polished form was administered to a larger group of participants to determine its final reliability and validity.

Data Analysis

To determine the psychometric properties of the *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition) subtests, the following statistical techniques were used: item difficulty index, item discrimination index, KR 20, Exploratory Factor Analysis and Confirmatory Factor Analysis.

Results

Contents of the preliminary and polished forms

The preliminary forms of *Pagkilala sa Salita* (Word Recognition) and *Pagbabaybay* (Spelling) consisted of 70 and 59 items respectively. Data from the participants who took the two forms were subjected to item analysis to determine item difficulty and discrimination. This served as the initial screening for the items, the purpose of which was to have a mixed pool of items for the polished form with at least a good level for both discrimination and difficulty. Table 2 presents the discrimination indices of the two subtests. For the items to be included in the polished form, they need to have a discrimination index of at

least .20 for *Pagkilala sa Salita* (Word Recognition) while the minimum index for *Pagbabaybay* (Spelling) was .30. The difference in the cut-off scores is due to the different nature of the two subtests.

Table 2
Item discrimination of Pagkilala sa Salita (Word Recognition) and Pagbabaybay

Item Number	Discrimination Indices	Description
Pagkilala sa Salita (Word Recognition)		
15 items: 5, 7, 13, 14, 18, 27, 29, 38, 44, 50, 51, 52, 54, 63, 65	.40 and above	Very Good
20 items: 2, 4, 8, 9, 10, 12, 15, 16, 25, 26, 32, 37, 39, 40, 41, 42, 47, 49, 60, 66	.30 to .39	Reasonably Good
25 items: 3, 6, 11, 17, 19, 20, 22, 28, 30, 31, 33, 34, 36, 43, 45, 46, 48, 53, 55, 56, 59, 64, 67, 68, 69	.20 to .29	Marginal Item
10 items: 1, 21, 23, 24, 35, 57, 58, 61, 62, 70	.19 and below	Poor
Pagbabaybay (Spelling)		
41 items: 1, 6, 9, 10, 11, 12, 13, 15, 17, 18, 19, 20, 21, 22, 23, 25, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 47, 48, 52, 53, 54, 56, 57, 58	.40 and above	Very Good
8 items: 2, 7, 14, 26, 27, 46, 49, 55	.30 to .39	Reasonably Good
6 items: 3, 16, 24, 44, 50, 51	.20 to .29	Marginal Item
4 items: 4, 5, 8, 59	.19 and below	Poor

Item difficulty was also computed. The results represent the intricacy of each of the test items as determined by their difficulty indices where the higher the index, the easier an item is. Similar to discrimination, the purpose is to have a mix pool of items that is not very easy and very hard. Table 3 contains the difficulty indices of both tests and for the items to be included in the

polished form, they need to have a difficulty index of at least .50 for *Pagkilala sa Salita* while the minimum index for *Pagbabaybay* (Spelling) was .20. Similar with the discrimination index, the difference in the cut-off scores is due to the different nature of the two tests.

Table 3
Item difficulty of Pagkilala sa Salita (Word Recognition)

Item Number	Difficulty Indices	Description
Pagkilala sa Salita (Word Recognition)		
9 items: 1, 21, 23, 24, 35, 57, 61, 62, 70	.91 and above	Very Easy
46 items: 2, 3, 4, 6, 8, 9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 22, 25, 26, 28, 30, 31, 32, 33, 34, 36, 37, 39, 40, 41, 42, 43, 45, 46, 47, 48, 53, 55, 56, 58, 59, 60, 64, 66, 67, 68, 69	.81 to .90	Easy
15 items: 5, 7, 13, 14, 27, 29, 38, 44, 49, 50, 51, 52, 54, 63, 65	.46 to .80	Optimum
No item	.30 to .45	Hard
No item	.29 downward	Very Hard
Pagbabaybay (Spelling)		
1 item: 5	.91 and above	Very Easy
10 items: 2, 9, 14, 16, 23, 24, 36, 49, 55, 59	.81 to .90	Easy
33 items: 1, 6, 7, 10, 11, 15, 18, 20, 21, 25, 26, 27, 29, 30, 32, 33, 34, 35, 38, 40, 41, 43, 44, 45, 46, 47, 48, 50, 51, 53, 56, 57, 58	.46 to .80	Optimum
12 items: 12, 13, 17, 19, 22, 28, 31, 37, 39, 42, 52, 54	.30 to .45	Hard
3 items: 3, 4, 8	.29 downward	Very Hard

Exploratory factor analysis

EFA was used to determine the factorial validity of the two subtests by examining whether the items would load in their pre-determined factors. The EFA results in Table 4 (Eigenvalues) and Figure 1 (Scree plot) supported the two hypothesized factors (*Pagkilala sa Salita* and *Pagbabaybay*).

Table 4
Eigenvalues of Pagkilala sa Salita (Word Recognition) and Pagbabaybay (Spelling)

Subscales	Eigenvalues
Pagkilala sa Salita (Word Recognition)	36.62186
Pagbabaybay (Spelling)	5.61261

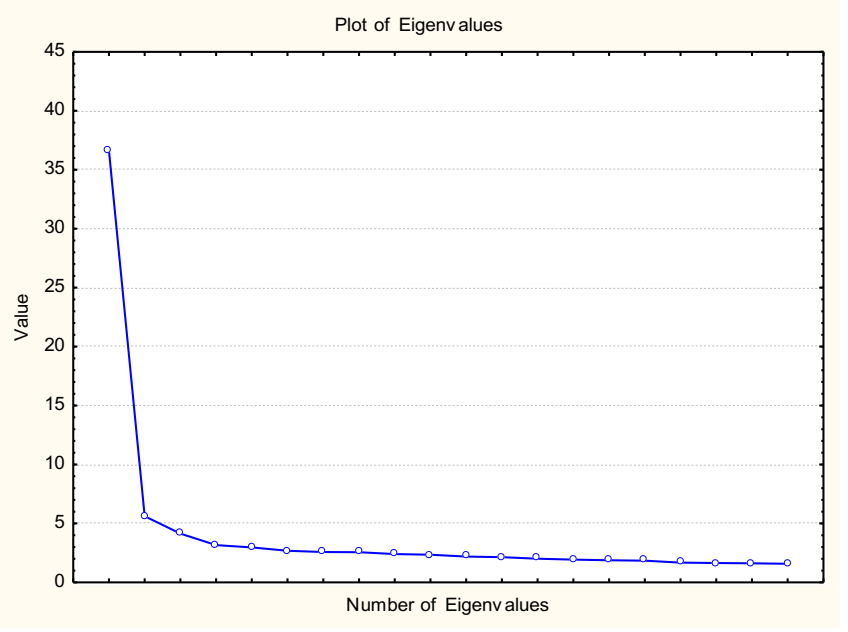


Figure 1. Scree plot of *Pagkilala sa Salita* (Word Recognition) and *Pagbabaybay* (Spelling).

Final composition of the polished form

As previously mentioned, an item can only be part of the polished form if it has a factor loading of at least .30 and acceptable difficulty and discrimination indices. Based on these criteria, only 27 of 70 items (38.57%) in the *Pagkilala sa Salita* (Word Recognition) subtest met the criteria. On other hand, in the *Pagbabaybay* (Spelling) subtest, 28 out of 59 items (47.46%) met the criteria (See Table 5).

Table 5
Difficulty, discrimination indices and factor loadings of the polished form

Item No.	Difficulty index	Discrimination Index	Factor Load
Pagkilala sa Salita (Word Recogntion)			
3	0.872	0.255	0.521
17	0.872	0.255	0.674
18	0.802	0.395	0.396
19	0.860	0.279	0.590
20	0.872	0.744	0.729
22	0.895	0.209	0.760
25	0.837	0.325	0.761
28	0.872	0.255	0.402
30	0.895	0.209	0.676
31	0.895	0.209	0.802
33	0.883	0.232	0.642
34	0.895	0.209	0.748
36	0.883	0.232	0.681
37	0.837	0.325	0.664
41	0.837	0.325	0.711
42	0.837	0.325	0.601
43	0.883	0.232	0.542
46	0.883	0.232	0.706
48	0.860	0.279	0.357
53	0.872	0.255	0.670
55	0.813	0.279	0.490
56	0.895	0.209	0.588

59	0.883	0.232	0.568
60	0.825	0.302	0.366
64	0.872	0.255	0.467
66	0.848	0.302	0.540
67	0.860	0.232	0.553
68	0.872	0.255	0.585
Pagbabaybay (Spelling)			
1	0.558	0.465	0.347
6	0.476	0.581	0.455
12	0.430	0.860	0.627
13	0.441	0.651	0.451
15	0.488	0.744	0.573
17	0.453	0.627	0.545
19	0.406	0.627	0.449
20	0.558	0.558	0.388
21	0.593	0.581	0.401
22	0.418	0.511	0.311
28	0.418	0.744	0.577
31	0.313	0.441	0.317
32	0.755	0.395	0.315
33	0.546	0.511	0.548
35	0.500	0.674	0.556
37	0.430	0.395	0.308
38	0.639	0.720	0.643
39	0.430	0.627	0.424
41	0.558	0.697	0.507
42	0.372	0.511	0.421
43	0.534	0.697	0.572
47	0.593	0.581	0.546
48	0.546	0.534	0.358
52	0.418	0.558	0.341
53	0.604	0.558	0.570
54	0.430	0.767	0.628
56	0.686	0.395	0.342

Reliability of the preliminary and polished forms

To check the reliability of the preliminary form of both *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition), KR 20 was used. Table 6 shows that both subtests are reliable. This suggests that the test items are sound, stable, and dependable. According to Friedenberg (1995), a test with high reliability is more favored over other types because it can be depended on to generate scores that are realistic estimates of the test taker's actual characteristics or knowledge.

Table 6
KR 20 reliability coefficients

Type of Test	Preliminary Form	Polished Form
Pagbabaybay (Spelling)	0.931	0.923
Pagkilala sa Salita (Word Recognition)	0.970	0.934

Confirmatory Factor Analysis

CFA was conducted to establish the final validity of the polished forms and confirm the factor structure of the two subtests. The following fit indices were used: Root Mean Square Error of Approximation (RMSEA), Population Gamma Index (PGI), Adjusted Population Gamma Index (AGPI), Joreskog GFI, Joreskog AGFI and the Chi-Square Model. The use of the said indices was based on the recommendations of Byrne (2010) and Hooper, Coughlan, and Mullen (2008). Results show that for both *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition), the chi square value is significant at 0.05 alpha, an indication of the departure of the data from the model. However, this may be due to the large sample size of the polished form (Anderson & Gerbing, 1998; Huang & Michael, 2000). The other fit indices indicate that the data from the *Pagbabaybay* (Spelling) met the recommended values for a good fitting level. On the other hand, the indices for the *Pagkilala sa Salita* (Word Recognition) did not meet the recommended values for a good fitting model but the obtained values can considered as adequate. Thus, the CFA results provided further support for the factorial validity of the two subtests.

Table 7
Goodness of fit indices

Goodness of fit index	Pagbabaybay (Spelling)	Pagkilala sa Salita (Word Recognition)	Recommended Values
RMSEA	0.038	0.079	0.06 and below
PGI	0.966	0.865	.90 and above
APGI	0.960	0.844	.95 and above
Joreskog GFI	0.947	0.849	.95 and above
Joreskog AGFI	0.938	0.825	.95 and above
Chi square	9609.68 ($p < .05$)	12601.39 ($p < .05$)	$p > .05$

Discussion

The purpose of the present study was to develop and validate the *Pagbabaybay* (Spelling) and *Pagkilala sa Salita* (Word Recognition) subtests of the Filipino Reading Achievement Test. The two domains were selected because spelling is believed to be a vital component of reading as it is one of the skill that supports it the most (Moats, 2006), while word recognition is essential since one's reading skill revolves around his ability to match the word he is reading with its corresponding orthographic-phonological representation that has previously been stored in his memory (Erickson et al., 2008). These two subtests, along with the previously published subtests on Vocabulary and Reading Comprehension (Cayubit, 2012), complete the battery of the Filipino Reading Achievement Test.

All items were written in Filipino since the researchers believe that this would be the best medium to assess Filipino children's reading ability. This stems from the notion that a Filipino child may be judged as poor in reading not because the student has impaired reading skills but because of limited command of English, the language used by most reading assessment tools.

In general, the analysis of the respondents' scores provides evidence that the two subtests have sound psychometric properties and that they have adequate validity and reliability. Validity evidences are from the results of content validation, exploratory factor analysis and confirmatory factory analysis. The multiple validity measures employed reflects the desire of the researchers to ensure that both subtests will indeed measure what they are supposed to

measure. In addition to being valid, the two subtests are also reliable, an indication that the scales can be dependent on to generate stable scores consistent with the domains being measured by over a period of time. These psychometric properties gives credence to the claim that the tests developed can potentially generate data that is reflective of the current reading ability of Filipino children with respect to their ability to spell and recognize words.

Considering its initial psychometric properties, stakeholders can now make use of the new tests to assess the reading skills of Filipino children. This is important because reading is a known predictor of achievement scores among students (O'Reilly & McNamara, 2007). Thus, it is expected that pupils who can read well are more likely to succeed or do well in school compared to those who have difficulty in reading. This is because reading has been found to be positively associated with behavioral engagement in school (Guo, Sun, Breit-Smith, Morrison, & Connor, 2015). Assessing reading skills is not only important for school but is also crucial in the future success and endeavors of children (Coker Jr., Jennings, Farley-Ripple, & MacArthur, 2018; Snow, Burns, & Griffin, 1998) because of its wide range implications for later academic attainment, economic success and other adult endeavors (Sullivan, Kohli, Farnsworth, Sadeh, & Jones, 2017) since almost all human activities would involve reading (Baddeley, Logie, & Nimmo-Smith, 1985).

Aside from this, the new tests can also be used to assess reading disabilities among Filipino children since the problem of reading appears to have existed among Filipinos for some time now (Luz, 2007). According to Luz (2007), many Filipino children seem to lack the ability to read and write and this is evident in the continuous decline in literacy rate of Filipinos. This means that many Filipinos appear to be incapable of reading and writing simple messages. However, Lutz (2007) also reported that literacy rate is often determined by census rather than assessment. This underscores the importance for a valid and reliable instrument that would accurately assess the reading skills and abilities of Filipinos. When the developed tests are used, the data from the children can serve as one basis for developing intervention programs that would address the reading problems experienced by Filipino children.

References

Alber, S. R., & Walshe, S. E. (2004). When to self-correct spelling words: A systematic replication. *Journal of Behavioral Education, 13,* 1-24. doi:10.1023/B:JOBE.0000011260.12674.a3

Anderson, J. C., & Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, *103*(3), 411-423.

Ariola, M. M. (2006). Principles and methods of research. Quezon City: Rex Printing Company, Inc.

Baddeley, A., Logie, R., & Nimmo-Smith, I. (1985). Component of fluency reading. *Journal of Memory and Language, 24,* 119-131.

Bai, Q. M. (2007). Study on reading strategies in college English teaching: Top-down or bottom-up? *Sino-US English Teaching, 4*(7), 27-31.

Boyd, D., & Bee H. (2015). *Life span development* (7th ed.). Essex, England: Pearson Education Limited.

Brady S. A. (1991). The role of working memory in reading disability. In S. A. Brady, D. P. Shankweiler (Eds). *Phonological processes in literacy, A tribute to Isabelle Y. Liberman* (p 129–152). Hills-dale, NJ: Lawrence Erlbaum.

Coker, Jr., D. L., Jennings, A. S., Farley-Ripple, E., & MacArthur, C. A. (2018). The type of writing instruction and practice matters: The direct and indirect effects of writing instruction and student practice on reading achievement. *Journal of Educational Psychology, 110*(4), 502-517.

Cayubit, R.F.O. (2012). Vocabulary and reading comprehension as a measure of reading skills of Filipino children. *The Assessment Handbook, 9,* 1-14.

De Dabat, A.V. (2006). Applying current approaches to the teaching of reading. *English Teaching Forum, 1*, 8-15.

Ehri, L. C. (2005). Development of sight word reading: Phases and findings. In M. J. Snowling & C. Hulme (Eds.). *The science of Reading: A handbook* (pp. 135-154). Oxford: Blackwell.

Erickson, K. A., Clendon, S. A., Cunningham, J. W., Spadorcia, S., Koppenhaver, D. A., Strum, J., & Yoder, D. E. (2008). Automatic word recognition: The validity of a universally accessible assessment task. *Augmentative and Alternative Communtication, 24*(1), 64-75.

Erten, I. H., & Karakas, M. (2007). Understanding the divergent influences of reading activities on the comprehension of short stories. *The Reading Matrix, 7*(3), 113-133.

Friedenberg, L. (1995). Psychological testing: Design, analysis and use. Massachusetts: Allyn & Bacon.

Friend, A., De Fries, J. C., Wadsworth, S. J., & Olson, R. K. (2007). Genetic and environmental influences on word recognition and spelling deficits as a function of age. *Behavior Genetics,* 37, 477-486.

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 4*, 1-40.

Gentry, J. R. (2007). A viewer's guide to assessing early literacy with Richard Gentry. Portsmouth, NH. Retrieved February 17, 2017, from http://www.heinemann.com/shared/onlineresources/E01044/viewers Guide.pdf

Gillet, J. W., Temple, C., & Crawford, A.N. (2008). *Understanding reading problems assessment & instruction* (7th ed.). USA: Pearson Education, Inc.

Graves, M. F., Wattes-Taffe, S. M., & Graves, B. B. (1999). *Essentials of elementary reading* (2nd ed.). Boston: Allyn & Bacon.

Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read.* East Sussex: Erlbaum.

Gunning, T. G. (2003). *Creating literacy instruction for all children* (4th ed.). USA: Pearson Education, Inc.

Guo, Y., Sun, S., Breit-Smith, A., Morrison, F. J., & Connor, C. M. (2015). Behavioral engagement and reading achievement in elementary school-age children: A longitudinal cross-lagged analysis. *Journal of Educational Psychology, 107*(2), 352-347.

Hayward, D. V., Stewart, G. E., Phillips, L. M., Norris, S. P., & Lovell, M. A. (2008). At-a-glance test review: Test of word reading efficiency (TOWRE). *Language, Phonological Awareness, and Reading Test Directory* (pp. 1-4). Edmonton, AB: Canadian Centre for Research on Literacy. Retrieved from http://www.uofaweb.ualberta.ca/elementaryed/ccrl.cfm.

Hook, P. E., & Jones, S. D. (2002). The importance of automaticity and fluency for efficient reading comprehension. *Perspectives, 28*(1) 9-14.

Huang, C., & Michael, W. B. (2000). A confirmatory factor analysis of scores on a Chinese version of an academic self-concept scale and its invariance across groups. *Educational and Psychological Measurement, 60,* 772-786.

Johnson, B. (2001). Toward a new qualification of nonexperimental quantitative research. *Educational Researcher, 30,* 3-13.

Karlin, R. (Ed.). (1973). *Reading for all.* Newark, DE: IRA.

Luz, J. M. (June, 2007). *A nation of nonreaders.* Retrieved from http://pcij.org/stories/a-nation-of-nonreaders/

Moats, L. C. (2006). How spelling supports reading: And why it is more regular and predictable than you think. *American Educator, 29*(4), 42-43.

Muter, V., Hulme, C., Snowling, M. J. & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading

development: Evidence from a longitudinal study. *Developmental Psychology*, *40*(5), 665-681.

O'Reilly, T., & McNamara, D.S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal, 44*(1), 161-196.

Phakiti, A. (2006). Theoretical and pedagogical issues in ESL/EFL teaching of strategic reading. *University of Sydney Papers in TESOL, 1,* 19-50.

Rajabi, P. (2009). Cultural orientation and reading comprehension model: The case of Iranian rural and urban students. *Novitas Royal, 3*(1), 75-82. Retrieved from http://www.novitasroyal.org/Vol_3_1/rajabi.pdf

Roe, B., Smith, S., & Burns, P. (2005). *Teaching reading in today's elementary schools.* (9th ed). New York: Houghton Mifflin Company.

Rumelhart, D. E. (1977). *Toward an interactive model of reading.* In S. Dornic (ed.), Attention and Performance IV. New York, NY: Academic Press.

Santrock, J. W. (2017). *Life span development* (16th ed.). New York, NY: McGraw Hill.

Scarborough, H. 2001. Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. Pp. 97-110 in S. B. Neuman & D. K. Dickinson (Eds.) *Handbook of Early Literacy.* NY: Guilford Press.

Snow, C., Burns, S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Strevens, P. (1977). *New orientations in the teaching of English.* London: Oxford University Press.

Sullivan, A. L., Kohli, N., Farnsworth, E. M., Sadeh, S., & Jones, L. (2017). Longitudinal models of reading achievement of students with learning disabilities and without disabilities. *School Psychology Quarterly, 32*(3), 336-349.

Tinker, M. A., & McCullough, C. M. (1975). *Teaching elementary reading.* Englewood Cliffs, N.J.: Prentice-Hall.

Torgeson, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed.

Treiman, R. (2001). Reading. *Blackwell Hanbook of Linguistics,* 664-672.

van Hell, J. G., Bosman, M. T., & Bartelings, M. (2003). Visual dictation improves spelling performance of three groups of Dutch students with spelling disorders. *Learning Disability Quarterly, 26,* 329-355.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*(2), 192-212.

Warrington, S. D. (2006). Building automaticity of word recognition for less proficient readers. *The Reading Matrix. 6*(1) 52-65

Ziegler, J. C., & Goswami U. (2005). Acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3-29.

Educational Measurement and Evaluation Review (2018), Vol. 9 Issue 1

Gender Differential Item Functioning in Polytomous Items: A Comparison of Three Methods

Consuelo T. Chua
Jose Q. Pedrajita
Kevin Carl P. Santos
University of the Philippines – Diliman

Abstract

The present study compared the consistency of the results of three non-parametric differential item functioning (DIF) techniques – the Cumulative Common Log-Odds Ratio (CCLOR), Standardized Mean Difference (SMD), and the Mantel Test (Mantel) in detecting gender DIF in the *Emotional Quotient Scale* – College Version. The sample comprised 1,229 college students (male = 657; women= 572) from a state university in the Philippines. The agreement of the DIF methods was determined using classification consistency and matching percentages. Results show that CCLOR and Mantel agreed perfectly in detecting gender DIF items. SMD, on the other hand, had a moderate to high agreement with the two other DIF techniques. The agreement among the DIF methods was lower when DIF effect size was considered.

Keywords: DIF, Classification Consistency, Standardized Mean Difference, Common Log-odds Ratio, Mantel Test

Introduction

The integrity of a test depends largely on the quality of its items. A test is valid when it contains items that measure relevant characteristics. On the contrary, when test scores depend on extraneous factors such as group membership (e. g., gender, social status), test bias is said to be present. Bias refers to the presence of systematic error that distorts the outcomes of a test for a

particular group (Camilli & Shepard, 1994; Osterlind, 1983). Bias is a technical term that simply refers to "the consistent distortion of a statistic" (Osterlind, 1983, p. 10) and does not necessarily suggest test unfairness (Penfield & Camilli, 2007).

Statistical methods to detect potentially biased test items were first developed in the 1970s. However, it was only in the 1980s that a 'general statistical framework,' now termed as differential item functioning (DIF), was developed as basis for the analysis of 'item statistical bias' (Penfield & Camilli, 2007, p. 126). The term 'potentially biased' is used to denote that DIF items are not automatically considered biased until the source of differential item functioning is explained.

DIF pertains to the difference in the performance of two or more matched groups (e.g., gender groups, age groups) on a test item. DIF occurs when members of groups who are similar in ability have different chances of obtaining a correct response or score on an item, leading to an unfair advantage for one group over the others (Penfield & Camilli, 2007; Roussos & Stout, 2004). For instance, if women are more likely to obtain a higher score on an essay item compared to men with similar proficiency then gender-based DIF is present. Although several groups may be involved in DIF analysis, two groups are normally compared - the reference group and the focal group. The reference group is the group that the test expects to favor while the focal group is the group that is likely to be disadvantaged by the test.

For dichotomous items, DIF pertains to the difference in the probability of a correct response between two groups with similar proficiency (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005). On the other hand, DIF is present in polytomous items when the probability of obtaining an item score differs between two matched groups. At present, there are numerous methods that are available for analyzing DIF for dichotomous items including item response theory (IRT) approaches, proportion-difference approaches, and common-odds ratio approaches (see Penfield & Camilli, 2007). Studies about DIF for dichotomous items are also abundant (e.g., Wiberg, 2009; Pedrajita, 2007; Gibson & Harvey, 2003). However, many tests contain items that are polytomously scored. Compared with dichotomous DIF, polytomous DIF deals with several score levels which makes DIF analysis more complex (Penfield & Camilli, 2007; Potenza & Dorans, 1995).

A number of methods have been proposed for polytomous DIF; several of which are extensions of procedures for dichotomous DIF including mean-difference approaches, multivariate hypergeometric distribution approaches,

and common odds ratio approaches (see Penfield & Camilli, 2007). However the agreement among several techniques for polytomous DIF has not been explored. The standardized mean difference (SMD) for instance is a non-parametric DIF technique that is used by the National Assessment of Educational Progress (NAEP) as an effect size estimator. On the other hand, the cumulative common log-odds ratio (CCLOR) is another polytomous DIF method that is also widely used to determine DIF effect size. The agreement between these two methods has yet to be examined.

Comparing the effect sizes of polytomous DIF techniques is important because DIF outcomes should be interpreted not only in terms of statistical significance but also based on the severity or size of potential bias. This is true because significant DIF outcomes may be observed "even for negligible departures from the null hypothesis" and small DIF effect sizes do not normally possess "practical value" (Meyer, Huynh, & Seaman, 2004, p. 335). Another DIF technique that is widely used is the Mantel Test (Mantel, 1963). This method is an extension of the Mantel-Haenszel procedure that has been widely used in previous studies. Thus, it would be beneficial to compare the Mantel test with SMD and CCLOR. In addition, these three methods are non-parametric statistics that are easier to understand compared to more complex DIF techniques such as IRT-based methods. Furthermore, parametric DIF techniques require that certain conditions are met (e.g., sample size) which limit their use in practical settings (Penfield, Glacobbi, & Myers, 2007). In contrast, the Mantel test, SMD, and CCLOR are non-parametric tests that do not require stringent assumptions. For this reason, these methods can be easily employed in many actual test situations. However, the performance of these methods in detecting DIF items has not been investigated empirically in the literature. This study hopes to fill in this gap by comparing the ability of these three methods in identifying DIF items.

The Mantel, SMD, and CCLOR tests have been applied in several studies across various tests to measure different types of DIF. For instance, CCLOR was applied by Penfield, Giacobbi, and Myers (2007) to detect gender DIF in the Exercise Imagery Inventory using total score of the relevant subscales as matching variable. Two out of the 19 items in the scale, one with moderate DIF and another with large DIF, were flagged as functioning differently between genders. In another study, standardized CCLOR was used as DIF effect size estimator for Mantel-Haenzel, alongside Ordinal Logistic Regression to identify cross-cultural DIF in the Student Questionnaire of the Program for International Student. The results yielded 14 items that were

similarly flagged as having either medium or large DIF using both procedures (Padilla, Baena, Hidalgo, & Sireci, 2011). On the other hand, the study of Wetzel and Hell (n.d.) examined gender DIF on the Allgemeiner Interessen-Struktur-Test [General Interest Structure Test] using CCLOR and IRT. DIF analyses show that the two procedures generally agree in flagging DIF items. An examination of the foregoing studies shows that CCLOR can be effectively used to identify DIF in scaled items, especially when DIF effect size is sought. CCLOR also has a good agreement with other DIF methods.

Similar with CCLOR, SMD has also been used to measure DIF in a variety of tests. Fletcher (2008) used SMD to measure the DIF effect size of the Likelihood Ratio Test in identifying ethnicity–based DIF on the Risk Perception Survey for Mellitus (RPS-DM). The outcomes of the analysis showed that the RPS-DM contained five items with strong DIF and one item with weak DIF. Similarly, Schwarz, Rich, and Podrabsky (2003) applied SMD together with the Linn-Harnisch procedure to examine DIF based on mode of test administration (on-line or paper and pencil). Two instruments were used for the analysis including an aptitude test and knowledge test on Reading, Mathematics and Language. A few items were flagged as having small to moderate DIF, some against the on-line group and others against the paper and pencil group. In another study, SMD was also used together with the Mantel test to identify gender DIF on an ordinal self-concept scale. Results showed that 42% of the items contained DIF which vary in both direction and magnitude (Young & Sudweeks, 2005). The aforementioned studies demonstrated that SMD can be used to measure DIF across different test types (e.g., cognitive and affective) and among varied DIF comparisons (e.g., gender, ethnicity, and test administration).

Aside from the previously mentioned study of Young and Sudweeks (2005), the Mantel procedure has also been applied to detect DIF in other contexts. For example, Henderson (2001) used Mantel together the Mantel-Haenszel (MH), SIBTEST, and Poly-SIBTEST to determine gender DIF on an academic high school exit test. The matching variable involved the corresponding total test score for each item. DIF results revealed that about 15% of the dichotomous items had DIF, while more DIF items were detected among polytomous items. The Mantel also displayed good agreement with MH in detecting DIF items. In another study, Cameron, Scott, Adler, and Reid (2014) identified age and gender-DIF on the Hospital Anxiety Depression Scale using the Mantel, ordinal logistic regression, and Rasch Analysis. The three methods similarly flagged three age-related DIF items, showing the general

consistency among the methods in detecting DIF. An inspection of the abovementioned studies illustrates that the Mantel is sensitive in detecting various forms of DIF and has high consistency with other DIF techniques.

Considering the aforementioned studies on CCLOR, SMD, and Mantel, it is worthwhile to compare the results of the three tests because SMD and CCLOR are both established effect size estimators that may be effectively used in combination with Mantel. Conducting this comparison would provide both statistical and practical interpretations of DIF, which allows for a more accurate interpretation of DIF results. In addition, the three methods have similar characteristics, including their applicability to ordinal data and the use of sum of scores as matching variable that facilitates DIF comparison on a same given test. Therefore, the purpose of this study was to compare the agreement or consistency among three polytomous DIF statistics – the SMD, MANTEL, and CCLOR in detecting gender DIF items in a polytomous emotional quotient test. The study specifically aimed to: (1) examine the consistency among the methods in detecting gender DIF based on statistical significance; and (2) determine the consistency between CCLOR and SMD in classifying DIF items based on substantial significance and effect size. Comparison based on substantial DIF and effect size was only possible for CCLOR and SMD because MANTEL by itself does not produce a measure of DIF effect size.

Method

Participants

The sample comprised 1,229 undergraduate students (657 are males; 572 are females) from the University of the Philippines - Diliman. The mean age of the respondents was 20. The respondents were mostly third to fifth year college students coming from varied science-related courses. Only respondents who were at least in third year college were selected because there were items in the scale that may not be applicable to younger students.

Research Instrument

The DIF analysis was performed using the *Emotional Quotient Scale* – college version *(EQS-C)* which was developed by Marquez (2002) and distributed by MAVEC Specialist Foundation Inc. An emotional quotient scale was chosen because studies have consistently shown differences between

genders on the construct 'emotional intelligence' (e.g., Tapia & Marsh, 2006; Rooy, Alonso, & Viswesvaran, 2005). It would therefore be significant to determine if gender DIF has a contribution to such measurement differences.

The *EQS-C* is a 140-item emotional intelligence test for college students and a modification of the *EQS* test for adults and employees (*EQS-AE)*. The scale has five response options ranging from *very true of me* to *very untrue of me*. The test has ten subscales: adaptability, communication (15 items), confidence (20 items), decision-making (15 items), empathy (15 items), interpersonal skills (23 items), motivation (7 items), innovation (6 items), teamwork (21 items), and trustworthiness (3 items). However, three subscales - innovation, motivation, and trustworthiness were excluded from the analysis because their Cronbach's alpha estimates were lower than r=.70. All the other subscales have moderate to high Cronbach's alpha (r=.75 to r=.86).

Data Collection Procedures

After obtaining the necessary permits from the concerned officials and faculty members of the University, the *EQS-C* was administered to 55 separate classes of students. The following procedures were followed for each data collection session. First, the students were requested to participate in the study by answering the *EQS-C*. They were also informed that participation was voluntary, and those who agreed to participate were requested to accomplish consent forms. Next, the students were provided with the test materials and were given directions on answering the test. Finally, the students answered the *EQS-C* and submitted their answer sheets. Each data collection session lasted for 30 minutes on the average. Data were collected during the first quarter of 2013.

Data Analysis

The initial step of data analysis involved detecting the gender DIF items in the *EQS-C* using CCLOR, MANTEL, and SMD. This procedure served as the basis for determining the agreement among the three DIF methods. Gender DIF detection for all three methods was implemented using a procedure similar to the one outlined by Penfield and Camilli (2007). The first step involved establishing the sum of scores for each subscale as the matching variable. The next step involved establishing the reliability of each matching variable by

computing for the Cronbach's alpha coefficient for each subscale. The third procedure concerned conducting a t-test to ensure that the scores of the two gender groups on the different subscales were matched. The fourth step concerned stratifying the respondents into three equally-spaced ability levels based on their total score per subscale. The final step involved applying three statistical techniques - CCLOR (Penfield & Algina, 2003), SMD, and MANTEL (Mantel, 1963) to detect gender DIF on each subscale or dimension of the test. Separate DIF analyses were conducted for each subscale of the *EQS-C* because DIF assessment assumes that all items "measure the same dimension." Therefore, in multidimensional tests such as the *EQS-C*, it is essential to decompose scores "into more homogenous subscores" to ensure the validity of the DIF outcomes (Dorans & Holland, 1993 as cited in Potenza & Dorans, 1995, p. 32). An item is flagged as DIF based on statistical significance if the DIF tests were significant at $\alpha = .05$. On the other hand, an item was flagged as having *substantial DIF* when the DIF effect size is at least moderate or large.

Cumulative Common Log-odds Ratio(CCLOR). The cumulative common log-odds ratio determines the difference between the focal and reference groups in terms of the odds of exceeding each category of the studied item, while controlling for target trait (Penfield, Giacobbi, & Myers, 2007). The cumulative common odds ratio was proposed by Liu & Agresti (1996) and applied to the detection of DIF of polytomous items by Penfield & Algina (2003).

CCLOR is computed based on the average estimated odds ratio of all the response categories of the item within each ability level. The average of the odds ratio is called the *cumulative common odds ratio estimator*. This estimator cannot have negative values and as such, its natural logarithm called the *cumulative common log-odds estimator* is taken to allow both negative and positive values (Penfield, Giacobbi, & Myers, 2007; Penfield & Algina, 2003).

The null hypothesis is that the odds of exceeding each response level of an item is the same for the two groups (Penfield, Giacobbi, & Myers, 2007). A zero *CCLOR* value indicates the absence of DIF; while values of *CCLOR* < 0 shows DIF in favor of focal group (males) and values of *CCLOR* > 0 shows DIF in favor of reference group (females). The DIF effect size of *CCLOR* was determined based on a classification scheme proposed by Penfield (2007), as shown below.

- Category AA (small) – when either *CCLOR* is not significantly different from zero or $|CCLOR| < .43$

- Category BB (moderate) – when *CCLOR* is significantly different from zero and $|CCLOR| \geq .43$ and either $|CCLOR| < .64$ or $|CCLOR|$ is not significantly greater than .43
- Category CC (large) - when $|CCLOR|$ is significantly greater than 0.43 and $|CCLOR| \geq .64$

The DIFAS 5.0 program, developed by Randall Penfield was used to compute for the cumulative common log-odds ratio, *CCLOR*; the standard error of *CCLOR*, and the standardized *CCLOR* (Z statistic). Corresponding p-values were computed using Excel in order to test the null hypothesis that *CCLOR* =0. A modified standardized *CCLOR* was also computed for each item in order to test whether the value of *CCLOR* is not significantly greater than .43. The formula used for computing the modified test statistic was [|CCLOR|-.43]/standard error.

The Mantel Test (MANTEL). The Mantel Test (Mantel, 1963), an extension of the Mantel-Haenszel technique is used to test the association between two matched groups on ordinal items (Welch & Hoover, 1993). It tests the null hypothesis of no association between group and the response variable.

An item is flagged as having DIF under the Mantel Test if the null hypothesis is rejected at α = .05. MANTEL by itself does not provide information on the direction of DIF and a classification scheme to categorize the effect size of DIF. The GMHDIF program developed by Fidalgo (2011) was used to detect DIF using this procedure.

Standardized Mean Difference (SMD). The standardized mean difference is a descriptive index that compares the item means of the focal and reference groups, after adjusting for differences in the distribution of members of the groups across the values of the matching variable (Zwick, Thayer & Mazzeo, 1997). The null hypothesis is that the population value of the standardized mean difference is zero. An SMD index of zero pertains to the absence of DIF. In general, positive SMD values correspond to DIF in favor of the reference group while negative SMD values correspond to DIF in favor of the focal group (Penfield, Giacobbi, & Myers, 2007). However, for this study, the SMD program was written such that positive SMD values pertain to DIF in favor of males (the focal group), whereas negative SMD values pertain to DIF in favor of females (the reference group). The R software was utilized to run the test using the SMD R Script provided in the study of Wood (2011).

The National Assessment of Educational Progress (NAEP) categorization scheme for classifying the effect size of DIF was used to interpret the size of DIF for standardized mean difference (John Donoghue, Personal Communication). The classification is shown below.

- Category AA (small or negligible) - if Mantel Chi-square p-value < 0.05 and |SMD/SD| ≤ 0.17
- Category BB (moderate) - if Mantel Chi-square is significant at p-value < 0.05 and |SMD/SD| > 0.17
- Category CC (large) – if Mantel Chi-square is significant at p-value < 0.05 and |SMD/SD| > 0.25

SMD refers to the standardized mean difference index while SD pertains to the group standard deviation of the item score. For the purposes of this study, only the value of SMD and not of Mantel was considered in determining the effect size of SMD. Items with significant SMD values at α =.05 were considered as DIF items and the value of |SMD/SD| as previously mentioned was used to classify DIF items.

Consistency Among Polytomous DIF Methods. The agreement among CCLOR, MANTEL, and SMD was determined using classification consistency and matching percentage. Classification consistency involves the simple procedure of comparing the number of DIF items consistently detected by the methods. On the other hand, matching percentages were computed by obtaining the ratio between the number of items detected using both procedures under comparison and the number of items detected using at least one procedure (Wiberg, 2009). The resulting matching percentages were interpreted as follows.

- High Matching Percentage - 75 to100%
- Moderate Matching Percentage - 50% to 74%
- Low Matching Percentage - less than 50 %

The consistency among the DIF methods was determined based on three DIF interpretations – DIF based on statistical significance, DIF based on substantial significance, and DIF based on effect size. The consistency of the three methods in detecting DIF items based on statistical significance was determined by considering all flagged DIF items regardless of effect size. On the other hand, the agreement between CCLOR and SMD to detect DIF based

on substantial significance was determined by considering only DIF items that have at least a moderate effect size or classified as category BB or CC. Finally, the consistency of the methods in flagging DIF based on effect size was determined by obtaining the frequency DIF items that were consistently classified as small, moderate, or large by the two procedures. Only CCLOR and SMD were considered in the DIF comparison based on substantial DIF and effect size because as previously mentioned, MANTEL by itself does not produce a measure of DIF effect size.

Results and Discussion

Gender DIF Items in the Emotional Quotient Scale

Cumulative Common Log-Odds Ratio (CCLOR). CCLOR detected 47 gender DIF items based on statistical significance in the *EQS-C*. Table 1 shows the gender DIF items that were detected for each subscale of the test and the corresponding DIF effect sizes. Twenty-four of these items were potentially biased toward males, which meant that males have a greater probability of attaining a higher score in these items compared to females with similar EQ levels. On the other hand, 23 DIF items were in favor of females. For these items, females have a higher chance of receiving a higher score on the items compared to males who have the same EQ level. However, out of the 47 gender DIF items, only 12 had substantial (moderate or large) amounts of DIF (Table 2). Four of these items had large DIF effect sizes, while eight had moderate DIF effect sizes. The rest of the DIF items only had small or negligible DIF which is not sufficient to conclude the presence of DIF.

Standardized Mean Difference (SMD). SMD flagged 45 gender DIF items in the seven subscales of the *EQS-C* (Table 1). Most of these items, 25 in all, were potentially biased towards males. For these items, males had a possible unfair advantage of obtaining higher scores compared to females with similar abilities on the corresponding subscale to which the items belong. On the other hand, 20 items were biased towards females which meant that females had a higher probability of obtaining higher scores in these items compared to males. Among the 45 DIF items detected, only 15 had substantial amounts of DIF (Table 2). Nine items had large DIF, while seven had moderate DIF. The rest of the items had small or negligible DIF.

The Mantel Test (MANTEL). The Mantel Test detected 47 gender DIF items (Table1). MANTEL does not provide a value to determine the direction of DIF – whether the item is in favor of the reference group or the focus group . Furthermore, MANTEL by itself does not provide an effect size estimate for the severity of DIF. Thus, when employing MANTEL for DIF detection, other DIF techniques that can serve as effect size estimators. Both CCLOR and SMD can serve this purpose.

All in all, the three DIF procedures detected 50 gender DIF items or 40% of the 124 items in the seven subscales of the Emotional Quotient Scale. However, only 16 out of the 50 items had substantial DIF or at least moderate DIF effect size. Further, of the 50 DIF items, 42 items were detected by all three methods. A greater number of DIF items (27 items) were found to be in favor of males compared to only 23 DIF items that were favorable to females. This is an unexpected result given that EQ scales mostly produce scores that are favorable to females (e.g., Rooy, Alonso, & Viswesvaran, 2005; Day & Carol, 2004).

The CCLOR and MANTEL procedures were equally sensitive in detecting DIF based on statistical significance and flagged 47 items each. The standardized mean difference procedure was slightly more conservative and detected only 45 DIF items. However, although CCLOR was more sensitive in flagging DIF items than SMD based on statistical significance, the former was a more conservative effect size estimator. For instance, items 6, 10, 18, and 98 all had moderate DIF under the SMD procedure but only had small DIF under the CCLOR procedure.

Table 1

Gender DIF Items in the Emotional Quotient Scale based on Statistical Significance

	CCLOR			SMD			MANTEL	
Item	CCLOR	p-value	In Favor of	SMD	p-value F	In Favor of	MANTEL	p-value
				Adaptability				
7	0.296	0.008	F	-0.125	0.008	F	7.027	0.008
18	0.376	0.001	F	-0.134	0.001	F	10.571	0.001
61	-0.326	0.003	M	0.132	0.003	M	8.688	0.003
90	-0.56	0.000	M	0.287	0.000	M	27.439	0.000
93	-0.275	0.016	M	0.099	0.014	M	5.871	0.015
125	0.316	0.010	F	-0.093	0.011	F	6.633	0.010
127	0.503	0.000	F	-0.187	0.000	F	18.810	0.000
128	0.346	0.002	F	-0.161	0.002	F	10.024	0.002
				Communication				
38	0.345	0.004	F	-0.140	0.004	F	8.339	0.004
47	0.439	0.003	F	-0.056	0.040	F	8.776	0.003
60	0.351	0.004	F	-0.119	0.005	F	8.080	0.005
78	-0.231	0.041	M	0.107	0.042	M	4.152	0.042
82	0.338	0.008	F	-0.097	0.007	F	7.168	0.007
84	NS	NS	NS	0.094	0.047	M	NS	NS
89	-0.305	0.007	M	0.134	0.009	M	7.159	0.008
94	0.252	0.040	F	-0.099	0.026	F	4.255	0.039
98	-0.402	0.000	M	0.174	0.000	M	12.109	0.001
				Confidence				
6	0.518	0.000	F	-0.161	0.000	F	18.502	0.000
9	-0.804	0.000	M	0.352	0.000	M	51.637	0.000
48	-0.332	0.003	M	0.131	0.003	M	8.660	0.003
49	0.721	0.000	F	-0.237	0.000	F	34.266	0.000
50	-0.238	0.028	M	0.112	0.028	M	4.861	0.028
95	0.586	0.000	F	-0.203	0.000	F	25.365	0.000
99	0.399	0.001	F	-0.133	0.001	F	11.772	0.001
113	-0.464	0.000	M	0.191	0.000	M	17.884	0.000
131	-0.36	0.002	M	0.133	0.002	M	9.666	0.002
134	0.326	0.004	F	-0.126	0.004	F	8.320	0.004
				Decision-Making				
73	-0.251	0.024	M	0.097	0.025	M	5.088	0.024
120	-0.424	0.000	M	0.188	0.000	M	14.741	0.000
123	0.356	0.004	F	-0.123	0.002	F	8.711	0.003
136	NS	NS	NS	0.081	0.049	M	NS	NS
				Empathy				
56	NS	NS	NS	0.124	0.046	M	NS	NS
1	0.294	0.013	F	-0.125	0.009	F	6.266	0.012
				Interpersonal Skills				
13	0.594	0.000	F	-0.341	0.000	F	27.502	0.000
32	-0.242	0.042	M	NS	NS	NS	4.127	0.042
54	-0.29	0.012	M	0.121	0.013	M	6.398	0.011
70	-0.589	0.000	M	0.215	0.000	M	19.945	0.000
76	0.65	0.000	F	-0.425	0.000	F	30.747	0.000
79	0.304	0.043	F	NS	NS	NS	4.111	0.043
114	-0.262	0.035	M	0.081	0.045	M	4.504	0.034
116	-0.3	0.020	M	0.087	0.016	M	5.381	0.020
135	-0.291	0.014	M	0.118	0.019	M	6.014	0.014
				Teamwork				
10	-0.4	0.000	M	0.178	0.000	M	12.675	0.000
31	-0.767	0.000	M	0.348	0.000	M	47.398	0.000
57	0.366	0.004	F	-0.094	0.001	F	8.202	0.004
58	-0.269	0.011	M	0.148	0.013	M	6.407	0.011
67	0.242	0.032	F	NS	NS	NS	4.516	0.034
104	0.305	0.015	F	NS	NS	NS	6.013	0.014
118	-0.351	0.004	M	NS	NS	NS	8.418	0.004
132	-0.283	0.027	M	0.072	0.025	M	4.889	0.027

Note: F=females; M=males; NS = Not Significant

Table 2
Items that have Substantial DIF using CCLOR and SMD

Item	CCLOR				SMD			
	CCLOR	*CCLOR* Modified	p-value (*CCLOR* ≥0.43)	Size of DIF	SMD	SD	SMD/ SD	Size of DIF
			Adaptability					
18	0.376	-0.466	0.679	Small	-0.134	0.774	-0.174	Mod
90	-0.560	1.215	0.112	Mod	0.287	1.032	0.278	Large
127	0.503	0.624	0.266	Mod	-0.187	0.728	-0.257	Large
			Communication					
47	0.439	0.060	0.476	Mod	-0.056	0.546	-0.103	Small
98	-0.402	-0.243	0.596	Small	0.174	0.954	0.183	Mod
			Confidence					
6	0.518	0.727	0.234	Small	-0.161	0.717	-0.224	Mod
9	-0.804	3.369	0.000	Large	0.352	0.954	0.369	Large
49	0.721	2.347	0.009	Large	-0.237	0.766	-0.309	Large
95	0.586	1.333	0.091	Mod	-0.203	0.810	-0.251	Large
113	-0.464	0.312	0.378	Mod	0.191	0.918	0.209	Mod
			Decision-making					
120	-0.424	-0.054	0.522	Small	0.188	1.003	0.188	Mod
			Interpersonal Skills					
13	0.594	1.426	0.077	Mod	-0.341	1.130	-0.302	Large
70	-0.589	1.205	0.114	Mod	0.215	0.857	0.251	Large
76	0.650	1.849	0.032	Large	-0.425	1.273	-0.334	Large
			Teamwork					
10	-0.400	-0.270	0.607	Small	0.178	0.933	0.190	Mod
31	-0.767	2.982	0.001	Large	0.348	0.927	0.375	Large

Note: DIF is substantial if test statistic is significant and DIF effect size is at least moderate; Mod=moderate; CCLOR Modified is the test statistic used to determine whether CCLOR is significantly greater than or equal to .43; SD = standard deviation

Detection Consistency among DIF Methods Based on Statistical Significance

The number of DIF items consistently detected by CCLOR, SMD, and MANTEL and the matching percentages of agreement among the methods based on statistical significance are shown in Table 3. MANTEL and CCLOR had a perfect agreement (100% matching percentage) in all subscales. The two methods commonly flagged a total of 47 items across the subscales, including items 7, 18, 61, 90, 93, 125, 127, 128, 38, 47, 60, 78, 82, 89, 94, 98, 6, 9, 48, 49,

50, 95, 99, 113, 131, 134, 73, 120, 123, 1, 13, 32, 54, 70, 76, 79, 114, 116, 135, 10, 31, 57, 58, 67, 104, 118, and 132 (Table 1). On the other hand, the comparison between CCLOR and SMD; and MANTEL and SMD both resulted to an agreement ranging from 50% to 100% which is moderate to high. CCLOR and SMD consistently detected all items except for items 84, 136, 56, 32, 79, 67, 104, and 118.

The same items were consistently detected by MANTEL and SMD. The perfect consistency between MANTEL and CCLOR shows that the two methods are compatible and can be used together to check the validity of DIF outcomes. The outcomes also suggest that caution should be used when interpreting items that are flagged as DIF by only one of the methods. Furthermore, the compatibility of the two methods provides a good indication that CCLOR can effectively serve as an effect size estimator for MANTEL. However, the use of SMD as effect size estimator for MANTEL is also appropriate especially if a more conservative DIF outcome is sought. The two methods can serve as check and balance for DIF test of significance.

In general, the comparisons between methods yielded average to high matching percentages, ranging from 50% to 100%. The agreement among all three methods was also generally high. All three methods consistently detected 42 out of the 50 DIF items in the seven subscales including items 7, 18, 61, 90, 93, 125, 127, 128, 38, 47, 60, 78, 82, 89, 94, 98, 6, 9, 48, 49, 50, 95, 99, 113, 131, 134, 73, 120, 123, 1, 13, 54, 70, 76, 114, 116, 135, 10, 31, 57, 58, and 132.

Table 3

Classification Consistency and Matching Percentage of the Three Methods in Detecting DIF Items based on Statistical Significance

	Adaptability		
	CCLOR	SMD	Mantel
CCLOR	8		
SMD	8 (100%)	8	
Mantel	8 (100%)	8 (100%)	8
	Communication		
	CCLOR	SMD	Mantel
CCLOR	8		
SMD	8 (89%)	9	
Mantel	8 (100%)	8 (89%)	8
	Confidence		
	CCLOR	SMD	Mantel
CCLOR	10		
SMD	10 (100%)	10 (100%)	
Mantel	10 (100%)	10 (100%)	10
	Decision-Making		
	CCLOR	SMD	Mantel
CCLOR	3		
SMD	3 (75%)	4	
Mantel	3 (100%)	3 (75%)	3
	Empathy		
	CCLOR	SMD	Mantel
CCLOR	1		
SMD	1 (50%)	2	
Mantel	1 (100%)	1 (50%)	1
	Interpersonal Skills		
	CCLOR	SMD	Mantel
CCLOR	9		
SMD	7 (78%)	7	
Mantel	9 (100%)	7 (78%)	9
	Teamwork		
	CCLOR	SMD	Mantel
CCLOR	8		
SMD	5 (62.50%)	5	
Mantel	8 (100%)	5 (63%)	8

Note: DIF items include all items that were flagged based on significant p-values

Consistency between CCLOR and SMD in Detecting Items with Substantial DIF

The classification consistency and matching percentages between SMD and CCLOR in detecting items with substantial DIF are presented in Table 4. As shown, the matching percentages between SMD and CCLOR varied according to subscale, which ranged from 0 to 100 %. The two methods consistently detected 10 out of the 16 items with substantial DIF (62.50% agreement) in all subscales combined, including items 90, 127, 9, 49, 95, 113, 13, 70, 76, and 31.

Higher matching percentages were found in subscales with more substantial DIF items such as confidence and adaptability. On the other hand, 0% matching percentages were found in subscales such as communication and decision-making, wherein only one item with substantial DIF was detected. Logically, if the two DIF procedures did not agree on the single substantial DIF item, then the matching percentage would automatically be zero.

The consistency between CCLOR and SMD was lower when substantial DIF was considered compared to when DIF was detected based only on tests of significance. This outcome is expected given that another criterion, DIF effect size, is also considered in the comparison. This means that even if the methods agree in terms of flagging DIF based on statistical significance, the severity of DIF detected does not necessarily coincide.

The outcomes show that the agreement among methods is not always perfect especially when substantial DIF is considered. This outcome emphasizes even more the importance of using more than one method in detecting DIF in order to validate the accuracy of DIF results especially when the size of DIF outcome is small. Items should only be interpreted as DIF when these are flagged by both methods.

Table 4
Classification Consistency and Matching Percentage of SMD and CCLOR in Detecting Items with Substantial DIF

	Adaptability	
	CCLOR	SMD
CCLOR	2	
SMD	2 (66.67%)	4
	Communication	
	CCLOR	SMD
CCLOR	1	
SMD	0 (0%)	1
	Confidence	
	CCLOR	SMD
CCLOR	5	
SMD	5 (100%)	5
	Decision-making	
	CCLOR	SMD
CCLOR	0	
SMD	0 (0%)	1
	Interpersonal Skills	
	CCLOR	SMD
CCLOR	3	
SMD	3 (100%)	3
	Teamwork	
	CCLOR	SMD
CCLOR	1	
SMD	1 (50%)	2

Note: DIF is substantial if test statistic is significant at α=.05 and DIF effect size is moderate or large

Consistency between CCLOR and SMD in Classifying Statistically Significant DIF Items based on Effect Size

The consistency between cumulative common log odds ratio and standardized mean difference in classifying statistically significant DIF items based on the effect size of DIF is shown in Table 5. Across all subscales, the two DIF methods had a higher consistency in classifying small DIF compared to moderate and large DIF effect sizes. The two methods consistently classified 26 out of the 39 items (64.10%) with small DIF using either procedure, including items 7, 61, 93, 125, 128, 38, 60, 78, 82, 89, 94, 48, 50, 99, 131, 134, 73, 123, 1, 54, 114, 116, 135, 57, 58, and 132.

CCLOR and SMD had a moderate level of consistency in classifying items with large DIF. The two methods consistently classified 4 out of the 8 items with large DIF (50% agreement), including items 9, 49, 76, and 31. The lowest agreement concerned classification for moderate DIF items. Here, the consistency between CCLOR and SMD was only 18.18% or 2 out of 11 moderate DIF items that were detected by either method. Only items 6 and 113 were consistently classified as moderate by the two methods. Some items that were classified as moderate by SMD were classified only as small by CCLOR. On the other hand, some items that were classified as large by SMD were only classified as moderate by CCLOR. Generally, CCLOR and SMD are not consistent in classifying the size or severity of DIF especially for small and moderate DIF.

Summary, Conclusions, and Recommendations

The present study compared the consistency among DIF methods in detecting potentially biased items in a polytomous scale. The detection consistencies of the methods based on statistical significance, substantial significance, and effect size classification were examined. In order to address the purpose of the study, the *Emotional Quotient Scale* was administered to college students. The outcomes of the test were subjected to DIF analysis using CCLOR, SMD, and MANTEL. The DIF agreement among the three methods was compared using classification consistency and matching percentages.

Table 5

Number of Items Consistently Classified by CCLOR and SMD based on DIF Effect Size

	SMD				
	Adaptability				
CCLOR	Small	Mod	Large	NDIF	Total
Small	4	2			6
Mod			2		2
Large					
NDIF					
Total	5	1	2	0	8
	Communication				
CCLOR	Small	Mod	Large	NDIF	Total
Small	6	1			7
Mod	1				1
Large					
NDIF	1				1
Total	8	1			9
	Confidence				
CCLOR	Small	Mod	Large	NDIF	Total
Small	5	1			6
Mod		1	1		2
Large			2		2
NDIF					
Total	5	2	3		10
	Decision-making				
CCLOR	Small	Mod	Large	NDIF	Total
Small	2	1			3
Mod					
Large					
NDIF	1				1
Total	3	1			4
	Empathy				
CCLOR	Small	Mod	Large	NDIF	Total
Small	1				1
Mod					
Large					
NDIF	1				1
Total	2				2
	Interpersonal Skills				
CCLOR	Small	Mod	Large	NDIF	Total
Small	4				4
Mod			2		2
Large			1		1
NDIF	2				2
Total	6		3		9
	Teamwork				
CCLOR	Small	Mod	Large	NDIF	Total
Small	3	1		3	7
Mod					
Large			1		1
NDIF					
Total	3	1	1	3	8

Note: NDIF = Not DIF

A total of 50 DIF items (40% of the 124 items) were detected by all three methods in the seven subscales of the *EQ Scale* that were included in the study. Of the 50 items, 42 items were flagged by all three methods. However, only 16 items had substantial DIF or an effect size of at least moderate. CCLOR and Mantel were equally sensitive in detecting DIF items and flagged 47 items each while SMD detected only 45 items.

The CCLOR, MANTEL, and SMD had moderate to high levels of consistency in detecting gender DIF. CCLOR and MANTEL had a perfect agreement in detecting gender DIF across all subscales that were investigated. On the other hand, the comparison between SMD and the two other methods yielded a moderate to high levels of consistency. The consistency between CCLOR and SMD in detecting DIF was lower when substantial DIF was considered. Finally, the classification consistency between CCLOR and SMD in classifying DIF based on effect size showed a higher agreement in classifying small DIF items compared to moderate and large DIF. Moderate DIF items were the least consistently classified by the two methods.

The present study provided several contributions for educators and test developers. Primarily, it explained the agreement among three non-parametric polytomous DIF techniques, not only in terms of statistical significance but also in terms of effect size classification. This is important because in reality, most data do not satisfy the conditions of parametric tests. Thus, a look into non-parametric DIF methods such as GMH, CCLOR, and SMD provide wider applications. Furthermore, a comparison of effect size measures between DIF techniques is helpful because interpretations of DIF that combine both statistical and practical significance provide a more accurate interpretation of item bias.

The study provides the following practical recommendations when performing DIF analysis. First, the study established that MANTEL and CCLOR yield very high or even perfect agreement in detecting DIF. Thus, these two procedures can be effectively used together when detecting DIF and classifying the size of DIF. Flagging an item using both procedures provides more confidence on the validity of the outcomes. Further, since user-friendly point-and-click programs such as GMHDIF and DIFAS are available, then the two procedures can be more practical to use compared to SMD which needs to be programmed in R or other software such as SAS. Penfield, Giacobbi, and Myers (2007) gave the same recommendation about the practicality of DIFAS for CCLOR. Further, since CCLOR, MANTEL, and SMD showed relatively high agreement in detecting DIF, caution should be made when interpreting

DIF items that are flagged by only one of these procedures. One suggestion is that an item should be flagged by at least two of the three procedures to be considered as DIF.

The study also provides suggestions for future research. The usual practice in DIF studies is to consider only moderate and large DIF items as biased. However, it has not been established yet whether or not large DIF items, by inspection, have more 'biased' content than items with small or moderate DIF. Are items with large DIF really more biased that small and moderate DIF items based on qualitative examination? It would be helpful to conduct studies that qualitatively compare the bias content of DIF items with different effect sizes. This way, one can determine whether it is justifiable to include only large DIF items when detecting DIF items.

References

Cameron, I. M., Scott, N. W., Adler, M., & Reid, I. C. (2014). A comparison of three methods of assessing differential item functioning (DIF) in the Hospital Anxiety Depression Scale: ordinal logistic regression, Rasch analysis and the Mantel chi-square procedure. *Qual Life Res, 23,* 2883-2888. doi: 10.1007/s11136-014-0719-3.

Camilli, G., & Shepard, L. (1994). *Methods for Identifying Biased Items.* Thousand Oaks:Sage Publications.

Day, A., & Carrol, S. (2004). Using an ability-based measure of emotional intelligence to predict individual performance, group performance, and group citizenship behaviors. *Personality and Individual Differences. 36*, 1443-1458.

Fidalgo, A. (2011). GMHDIF: A computer program for detecting DIF in dichotomous and polytomous items using Generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement, 35* (3), 247-249.

Fletcher, J. (2008). *Detecting Differential Item Functioning (DIF) in the Diabetes Risk Perception Survey.* (Doctoral Dissertation). Retrieved from https://fordham.bepress.com/dissertations/AAI3353768/.

Gibson, S. G., & Harvey, R. J. (2003). Gender and ethnicity based differential item Functioning on the armed services vocational aptitude battery. *Equality, Diversity, and Inclusion: An International Journal. 22*(4), 1-15.

Henderson, D. L. (2001). *Prevalence of gender DIF in mixed format high school exit examinations.* Retrieved from https://files.eric.ed.gov/fulltext/ED458284.pdf.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Liu, I-M., & Agresti, A. (1996). Mantel-Haenszel-Type inference for cumulative odds ratios with a Stratified ordinal response. *Biometrics, 52*(4), 1223-1234.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel Procedure. *Journal of the American Statistical Association, 58*, 690-700.

Marquez, A. T. (2002). *Emotional Quotient Scale Manual.* Quezon City: Mavec Specialists Foundation Inc.

Meyer, J., Huynh, H., & Seaman, M. (2004). Exact small-sample differential item functioning Methods for polytomous items with illustration based on attitude survey. *Journal of Educational Measurement, 41*, 331-344.

Osterlind, S. (1983). *Test Item Bias.* Beverly Hills: Sage Publications.

Padilla, J. L, Baena, I. B., Hidalgo, M. D., & Sireci, S. G. (October 2011). *Cognitive interviewing evidence on DIF in Polytomous Items of the Student Questionnaire of the PISA.* Paper presented at the 42th Annual Conference of the Northeastern Educational Research Association, Rocky Hill, USA. Retrieved from http://digibug.ugr.es/bitstream/handle/10481/24229/Padilla_Benitez_Hidalgo_Sireci_NERA2011.pdf;jsessionid=620721BEE5F1A5E9A5AF798A529E26CA?sequence=1

Pedrajita, J. O. (2007). *Item Bias Elimination Models for Test Validity and Reliability.* Unpublished doctoral dissertation, University of the Philippines, Diliman.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*(3), 335-355.

Penfield, R. D., & Algina, J. (2003). Applyling the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penfield, R. D., & Camilli G. (2007). Differential item functioning and item bias. In C. Rao & S. Sinharay (Eds.), *Handbook of Statistics Psychometrics* (Vol. 26, pp. 125-167). Amsterdam: Elsevier.

Penfield, R. D., Giacobbi, P. R., & Myers, N. D. (2007). Using the cumulative log-odds ratio to identify differential item functioning of rating scale

items in the exercise and sports sciences. *Research Quarterly for Exercise and Sport, 78(*5), 451–464.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomous scored items: a framework for classification and evaluation. *Applied Psychological Measurement. 19*(1), 23 – 37.

Rooy, D. L., Alonso, A., & Viswesvaran, C. (2005). Group differences in emotional intelligence scores: theoretical and practical implications. *Personality and Individual Differences.38*, 689-700.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis: Detecting DIF items and testing DIF hypotheses. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for Social Sciences* (pp. 107-115). Thousand Oaks: Sage.

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests.* Retrieved from https://files.eric.ed.gov/fulltext/ED477932.pdf.

Tapia, M., & Marsh, E. (2006). A validation of the emotional intelligence inventory. *Psicothema, 18*, 55-58.

Welch, C., & Hoover, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education*, *6*(1), 1-19.

Wetzel, E., & Hell, B. (n.d.). *Differential item functioning in the AIST-R.* [PDF Slides]. Retrieved from http://www.ecpa11.lu.lv/files/Wetzel.pdf.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three Methods using real data. *International Journal of Testing, 9*, 41-59.

Wood, S. W. (2011). *Differential item functioning procedures for polytomous items when examinee sample sizes are small.* (Doctoral dissertation). Retrieved from: http://ir.uiowa.edu/etd/1110.

Young, E. L., & Sudweeks, R. R. (2005). Gender differential item functioning in the Multidimensional Self Concept Scale with a sample of early adolescent students. *Measurement and Evaluation in Counseling and Development, 38*(1), 29-43.

Zwick, R., Thayer, D., & Mazzeo, J. (1997). *Describing and Categorizing DIF in Polytomous Items* (CRE Board Professional Report No. 93-10P & ETS Research Report 97-05). New Jersey: Educational Testing Service.

Educational Measurement and Evaluation Review (2018), Vol. 9 Issue 1

The Moderating Role of Defensive Pessimism in the Relationship Between Test Anxiety and Performance in a Licensure Examination

Rene M. Nob
De La Salle University Manila
Alyonna Marie L. Bumanglag
Genevie Mae A. Diwa
Guia Isabel Ponce
St. Paul University Manila Philippines

Abstract

This study aims to determine if the dimensions of test anxiety (worry and emotionality) can negatively predict test takers' performance in a licensure examination. It also aims to test if defensive pessimism can buffer these predictive relationships. The study involved data from 101 individuals who took the Philippine licensure examination for Occupational Therapy and Physical Therapy. Results from logistic regression analysis reveal that worry negatively predicts examination performance. However, emotionality turns out to be a positive predictor, after controlling for worry. Furthermore, defensive pessimism weakens the negative effect of worry on examination performance, but did not serve as a moderator in the relationship between emotionality and examination results. Future research directions and some practical implications are further discussed.

Keywords: test anxiety, worry, emotionality, defensive pessimism, licensure examination

Introduction

Assessment remains to be an integral part of our educational system. While schools and teachers are encouraged to make use of a variety of

methods of assessment, summative tests remain to be a popular choice. Some tests matter more than others. For example, passing or failing high-stakes tests, such as licensure examinations, is said to have more serious consequences for the individual (Cole & Osterlind, 2008). It is very important that performance in these exams reflect students' true abilities. However, other individual factors may also influence the outcomes of exams. For example, test anxiety has been identified in previous research to have detrimental effects on test performance (Wong, 2008). Other lines of research also suggest that some individuals harness their education-related anxieties to prepare for forthcoming academic challenges, such as in the case of preparing for examinations (Norem & Cantor, 1986). Such individuals are said to be defensive pessimists. To the best of our knowledge, no study has been conducted about the potential protective role of defensive pessimism against the detrimental effects of test anxiety. Hence, this study would like to investigate if defensive pessimism will buffer the impact of test anxiety on performance in a licensure examination.

High-Stakes Tests and Test Anxiety

High-stakes tests are said to be assessments with meaningful consequences to the students (Cole & Osterlind, 2008). Some tests serve as basis for students to accelerate to another grade level, while others serve as requirement for entering college. High-stakes tests, especially licensure examinations, ensure the public that the individuals who passed meet the minimum standards to perform their respective professions. Failure to pass a licensure examination may cause a major career setback for the individual, such as not being able to practice his or her profession.

Because of the seeming importance of licensure examinations, many examinees may feel anxious before or even during the test. Such experience is referred to as, text anxiety. Anxiety is a highly unpleasant affective state similar to intense fear, which can include feelings of threat, vague objectless fear, a state of uneasiness and tension, and a generalized feeling of apprehension that can affect an individual's concentration in various situations (Basavanna, 2000). Anxiety is characterized by high arousal, negative valence, uncertainty, and a low sense of control (Gray, 1991). It usually occurs when an anticipated event is expected to make demands for which a person is unprepared for and therefore, lacks the necessary coping skills (Costello, 1976). Anxiety has been associated to error-related negativity via reduced active goal maintenance and compensatory

reactive control efforts (Moser, Moran, Schroder, Donnellan, & Yeung, 2013). It can also be conceptualized as a state of distress that can affect an individual's performance in reaction to situations that resemble previous undesirable events. Individuals who feel anxious tend to focus on the potential negative outcomes, which they may have experienced before, and believe that those outcomes are more likely to happen again in similar situations (Lerner & Keltner, 2001). This is the reason why most anxious individuals tend to escape the potentially unfavorable event.

Test anxiety is a strong negative emotional reaction that students feel before and during an examination (Akca, 2011; Hong & Karstensson, 2002). This is likely because of students' fear of evaluation (Liebert & Morris, 1967). Such may be related to having an avoidance temperament (Liew, Lench, Kao, Yeh, & Kwok, 2014), and may develop among students who experienced failure, even after exerting effort (Liepmann, Marggraf, Felfe, & Hosemann, 1992). Test anxiety is said to be "directly related to fears of negative evaluation, dislike of tests, and less effective study skills" (Hembree, 1988, p. 73). This happens when exam takers feel threatened by a test they are about to take or are currently taking, which triggers certain negative reactions.

There are two components of test anxiety: worry and emotionality (Morris & Liebert, 1970; Damer & Melendres, 2011). Worry is defined as a cognitive expression of concern over an impending evaluative performance. This may include pessimistic expectations, thoughts about possible negative outcomes, self-criticism, overwhelming fear about failing grades, and absent-mindedness (Berk & Nanda, 2006; Zeidner, 1998). Students who predict failure in the exam would experience anxiety and may see the exam as threatening. As a result, students who are test-anxious are filled with thoughts of self-deprecation such as, "I am going to fail this examination." The focus of the individuals is on the consequences and implications of failure rather than the examination itself (Zeidner, 1998). Worrying causes the mind to slow down by suppressing clear thought, resulting to problem-solving processes becoming more complex (Akca, 2011).

Another dimension of test anxiety according to Morris and Liebert (1970) is emotionality. Emotionality can be considered as the physiological component of test anxiety. Individuals may experience tensed muscles, fast heart rate, the feeling of sickness, dizziness, sweating, and shaking when taking a test (McDonald, 2001). These physiological reactions are said to decrease the concentration of students, thus, resulting to poor performance in tests (Arguelles, McCraty, & Rees, 2003; McCraty, 2005).

Test Anxiety and Test Performance

Given the gravity of the consequences of passing or failing a high-stakes test such as a licensure examination, it is important that students' performance be reflective of their true potential as future practitioners. However, literature about test anxiety indicates that, such academic emotion may actually get in the way of students' optimum performance. For instance, Wong (2008) reported that there is a small to moderate negative relationship between test anxiety and academic performance. According to Eysenck and Calvo (1992), anxiety often weakens performance especially under test conditions. Highly anxious people who report to have greater worry and emotionality tend to perform poorly than students with low test anxiety. Wine (1971) suggested that the loss of focus during a task marks the difference between a high test-anxious person and low test-anxious person. A low test-anxious person focuses well on test-relevant stimuli while performing a task, but a high test-anxious person focuses on test-irrelevant stimuli. Splitting of attention because of irrelevant stimuli could interfere with the performance (Wine, 1971).

Looking at the cognitive and emotional components of test anxiety and their relationship with measures of academic performance, the study of Cassady and Johnson (2002) reveals that cognitive test anxiety (i.e., worry) is negatively associated with test scores in various course examinations. On the other hand, while excessive emotionality resulted to poor test performance, moderate levels of arousal seem to be beneficial (Cassady & Johnson, 2002).

Chapell et al. (2005) found that among undergraduate students, both worry and emotionality are negatively correlated with both grades and cumulative grade point average (CGPA). However, it should be noted that the correlation coefficient for emotionality and grades, and emotionality and CPGA were much less than the correlation coefficient between worry and grades, and worry and CGPA (Chapell, et al., 2005). Furthermore, among graduate student participants in the same study, emotionality was no longer associated with grades and CGPA (Chapell, et al., 2005).

In the study of Rana and Mahmood (2010), regression results reveal that worry negatively predicted students' academic achievement. However, upon controlling for worry, emotionality no longer served as a unique predictor. This is despite the fact that emotionality negatively correlated to achievement, based on the preliminary correlation analysis (Rana & Mahmood, 2010).

These findings suggest that worry is a more consistent negative correlate of measures of academic performance (Cassady & Johnson, 2002; Chapell, et

al., 2005; Rana & Mahmood, 2010). However, findings with regard to emotionality are less consistent. While emotionality is negatively correlated with academic performance (Chapell, et al., 2005; Rana & Mahmood, 2010), the literature suggests that it may not have a unique impact on achievement beyond the influence of worry. Emotionality may be mostly nested in worrying, and unlikely to have a negative impact outside that which worrying affords (Deffenbacher, 1977; Morris, Smith, Andrews, & Morris, 1975; Schwarzer, 1984). Furthermore, there is also some evidence that emotionality might even be optimal in moderate levels; in such a way that it might actually enhance performance in a test (Cassady & Johnson, 2002). Nevertheless, the result of the study of Cassady and Johnson (2002) implies that high level of emotionality is detrimental to performance. Hence, the present study proposed the following hypothesis: (H1) Worry will negatively predict examination performance. (H2) Emotionality will negatively predict examination performance to a lesser degree compared to worry.

Buffering Effect of Defensive Pessimism

Anxiety need not always has to have detrimental effects on test performance. There are some people who utilize anxiety as a means to motivate themselves to prepare for examinations (Norem & Cantor, 1986). Such people are those who possess defensive pessimism. According to the Theory of Self-Regulated Learning, students can become active participants in the learning process (Pintrich, 2004). They are potentially capable of being aware of the nature of the task, learning experiences, setbacks, and end goals, for the purpose of controlling these various aspects of learning, to achieve desired outcomes (Pintrich, 2004). The use of specific strategies is crucial in becoming a self-regulated learner. One strategy used to negotiate with undesirable emotions and enhance motivation is defensive pessimism (Garcia, 1995; Norem & Cantor, 1986; Pintrich, 2004).

Defensive pessimism is acknowledged as a self-regulated learning strategy that involves "setting unrealistically low expectations in a risky situation in an attempt to harness anxiety so that performance is unimpaired" (Norem & Cantor, 1986, p. 1208). An individual's pessimism is said to be defensive because there is an existing inconsistency between that individual's past successful performance and the low expectation they have for future success (Thompson & le Fevre, 1999). In order for these individuals to protect themselves from potential disappointment for their performance, they set low expectations for

themselves, or expect the worst out of the situation. In turn, they use the anxiety generated by these thoughts as motivation for them to perform better. According to Terada and Ura (2015), defensive pessimists can attain higher levels of performance because they control their negative thinking beforehand and prepare for the worst possible outcomes. Norem and Cantor (1986) demonstrated through an experiment that although defensive pessimists gave lower predictions in their performance and scored high on anxiety, they performed as well as optimists. This shows that individuals, depending on how they view the situation at hand, can moderate the usual effects of anxiety on their performance. Hence, it is hypothesized that: (H3) Defensive pessimism will buffer the negative impact of worry and emotionality on examination performance.

Method

Participants

The participants for this study were 101 exam takers who took the Physical Therapists (PT) Licensure Examination and Occupational Therapists (OT) Licensure Examination on the second month of last year. Out of 101 participants, 73 are females (72.3%). Their age ranged from 20 to 28 years old (M_{age}=22.6, SD=1.639). The participants came from 31 different schools and 2 review centers. There were 47 participant who attended in review center 1 (46.5%) and 54 attended review center 2 (53.5%).

Measures

Informed consent was secured from the participants prior to data gathering. It was emphasized in the consent form that the results of the licensure examination, whether they pass or fail, will also be used in the analysis. Weeks before the schedule of their licensure examination, two measures were administered to assess participants' worry, emotionality, and defensive pessimism.

Test Anxiety Inventory (TAI). The TAI was developed by Spielberger (1980) based on his extensive research. It is a 20-item self-report instrument of test anxiety, in which the respondents would rate a set of statements using a four-point scale: 1=almost never; 2 sometimes; 3=often; 4= almost always.

There is a total score for test anxiety, but there are subscales for the two components: Worry (TAI/W) and Emotionality (TAI/E). Eight items from the TAI is categorized as worry items (numbers 3, 4, 5, 6, 7, 14, 17, and 20); another eight items represent emotionality items (numbers 2, 8, 9, 10, 11, 15, 16, and 18). All of these aforementioned items, together with four remaining items (numbers 1, 12, 13 and 19) comprise the total test anxiety. It should be noted that item 1 is scored in reverse. In the current study both subscales showed good reliability. TAI/W obtained an alpha of .82, while TAI/E has an alpha of .90.

Defensive Pessimism Questionnaire (DPQ). The DPQ is a 17-item questionnaire designed by Norem (2001) to measure pessimism (items 1, 2, 6, and 15) and reflectivity (items 4, 7, 8, 10, 12, 14, 16 and 17) regarding possible outcomes. Each item is scored from 1 (not at all true of me) to 7 (very true of me). Items 2 and 16 are scored in reverse, while items 5 and 9 are fillers. Items 11 and 13 are said to be experimental items, and hence, was treated as fillers as well. Item 3 is used to differentiate defensive or realistic pessimists. As a measure of overall defensive pessimism, Norem (2001) instructs to sum the ratings for pessimism and reflexivity items. For the current research, the average ratings were computed, instead. In this study, the overall measure of defensive pessimism has an acceptable reliability, with an alpha of .72.

Examination Performance. Given that there is limited access to the actual scores of the participants in their respective licensure examinations, examination performance was simply coded as 0 for fail and 1 for pass.

Data Analysis

Pearson *r* correlation was used to initially analyze the associations among the different constructs. Given that the outcome variable, which is examination performance, was measured as a binary: pass or fail, while the predictors were measured as continuous variables, logistic regression was used to test the main hypotheses.

Results

This study intended to determine whether test anxiety can negatively predict the results of PT and OT licensure examinations and whether defensive pessimism can buffer such negative effects. It was hypothesized that the test

anxiety components, emotionality, and worry will negatively predict performance in the PT and OT licensure examination. Likewise, it was hypothesized that defensive pessimism, as a moderator, can buffer the relationship between the two components of test anxiety and licensure examination performance. In order to address the research problems, a series of logistic regression analysis was conducted.

Table 1 shows the mean and standard deviation values of the variables that were included in the study. Likewise, Pearson *r* correlations were computed to evaluate the relationships among the variables.

Table 1
Mean, Standard Deviations, and Zero-Order Correlations of the Licensure Exam Results, Emotionality, Worry, Test Anxiety, and Defensive Pessimism

	M	*SD*	2	3	4
(1) Licensure Exam			-.048	-.341***	-.165
(2) Emotionality	2.501	.6693	--	.740***	.271**
(3) Worry	2.225	.5772		--	.410***
(4) Def. Pessimism	5.027	.7376			--

* p<.05; **p<.01; ***p<.001

Correlation analysis reveals that worry and emotionality are strongly correlated (r =.740). Between the two, only worry is negatively associated with licensure exam performance to a moderate degree (r = -.341). Furthermore, defensive pessimism is also correlated with emotionality (r = .271) and worry (r =.410). The results are reasonable since the experience of defensive pessimism is based on the experience of anxiety over future performance.

One of the objectives of the study is to determine if worry and emotionality negatively predicts the results of the PT and OT licensure examinations. Table 2 shows that collectively, the predictor variables significantly accounts for 20.3% of the variance in the licensure examination result, χ^2 (3) = 22.923, p=.0001.

Specifically, worry negatively predicted (B= -2.977, p=.001), while emotionality positively predicted (B=1.834, p=.004) the result of the licensure examination. Defensive pessimism, on the other hand, did not serve as a significant predictor (B=.009, p=.981). This shows that those who experience worry are more likely to fail the examination, while those who experience emotionality are more likely to pass, assuming that the other predictor variables

were held constant. While the result involving worry supports one of our hypotheses, the result with regard to emotionality, as a component of test anxiety, did not. Though emotionality is considered as a component of test anxiety, it shows a positive, instead of a negative impact on examination performance. This means that, assuming that participants have the same level of worrying, those who experience more physiological arousal, such as nervousness, are actually more likely to pass the licensure exam.

Table 2
Results of Worry, Emotionality and Defensive Pessimism as Predictors of Licensure Exam Performance

Variables	*B*	*SE*	*p*
Worry	-2.977	.755	.001
Emotionality	1.834	.638	.004
Defensive Pessimism	.009	.366	.981

Cox & Snell R^2 = .203
Omnibus $\chi^2(3)=22.923$, $p=.0001$

The final goal of this study is to determine whether defensive pessimism can moderate the impact of worry and emotionality on performance. Separate analyses were made for worry and emotionality, as predictor variables. Since these two are highly correlated, we decided that when analysis was conducted for one of them, the other should serve as a covariate, and hence was controlled for. For example, when we analyzed if defensive pessimism moderates the effect of worry on licensure exam result, emotionality was controlled for by including it in the analysis as a covariate. This allowed looking at the unique pattern of relationships under the assumption that the participants have a constant level of emotionality.

Table 3
Result of the Moderation Analysis among Worry, Defensive Pessimism, and Exam Results

Variables	*B*	*SE*	*p*
Worry	-3.2906	.8305	.0001
Defensive Pessimism	-.0466	.4098	.9095
Worry x Def. Pessimism	1.3292	.7053	.0595
Emotionality (control)	2.0747	.6674	.0019

Cox & Snell R^2 = .2363, p=.0001

Result of the analysis of the moderating role of defensive pessimism on the relationship between worry and licensure examination reveals that the overall model accounts for 23.63 % of the variance in the exam results. It was revealed earlier that worry negatively predicts the performance in the test. However, a marginal significant interaction effect shown in Table 3 suggests that defensive pessimism somehow moderates the relationship between worry and exam results (B=1.3292, p=.0595). This means that while those who failed were more likely to have experienced more worry in general, such may also depend on whether that person resorts to defensive pessimism or not.

Table 4
Conditional Effects of Worry Depending on the Level of Defensive Pessimism

Level of Defensive Pessimism	Effect (*B*)	*SE*	*p*
High	-2.3101	.8427	.0061
Average	-3.2906	.8305	.0001
Low	-4.2710	1.1003	.0001

Table 4 shows the conditional effects of worry across low, average, and high levels of defensive pessimism. While the effect of worry remains to be negative across all the levels of defensive pessimism, it can be noticed that the impact diminishes at higher levels of defensive pessimism. This signifies that while defensive pessimism may not exactly buffer the negative effect of worrying, such negative effect is of lesser magnitude for defensive pessimists. That is, while worry can contribute to failing a PT and OT licensure examination, such is less likely when one utilizes anxiety to prepare for such

summative assessment. This result somehow provides support to our second hypothesis.

Table 5
Result of the Moderation Analysis among Emotionality, Defensive Pessimism, and Exam Results

Variables	*B*	*SE*	*p*
Emotionality	1.8411	.6532	.0048
Defensive Pessimism	.1080	.3927	.7834
Emo x Def. Pessimism	.6359	.6078	.2954
Worry (control)	-2.9590	.7904	.0002

Cox & Snell R^2 = .2125, p=.0001

With regard to emotionality, Table 5 reveals that the model accounts for 21.25% of the variance in the exam results. Moderation analysis clearly shows that there is no interaction effect (B= .6359, p=.2954). This means that defensive pessimism does not moderate the relationship between emotionality and licensure exam performance. Therefore, assuming the students have the same amount of worry experienced, those who report more physiological arousal are more likely to pass the licensure exam, regardless if they engage in defensive pessimism or not.

Discussion

The objective of the study was to see if test anxiety, particularly worry and emotionality, can negatively predict the result of the licensure examination in Physical Therapy and Occupational Therapy. Furthermore, the current investigation aimed to find out if defensive pessimism, as a self-regulated learning strategy, can buffer such anticipated negative effects of worry and emotionality on examination performance.

As expected, the two components of test anxiety are correlated, suggesting that they are dimensions of the same psychological construct. Furthermore the correlation was not too strong that they can be regarded as grossly overlapping.

While it was hypothesized that worry and emotionality will negatively predict performance in the licensure examination, results only provide partial supported such predictions. Worry indeed negatively predicted examination results. As argued earlier, excessive worrying about tests and its consequences may take away important resources from students (Zeidner, 1998), resulting to the possible impairment of various problem-solving processes (Akca, 2011). This finding is consistent with previous results (Cassady & Johnson, 2002; Chapell, et al., 2005; Rana & Mahmood, 2010); therefore, strengthening the theorized negative impact of worry on measures of achievement.

On the other hand, the hypothesized negative relationship between emotionality and examination performance not only failed to gain support, but went to the direction that is opposite of what was expected. Instead of a negative relationship, emotionality positively predicted results of the licensure exam. Previous findings with regard to the impact of emotionality have been inconsistent (Cassady & Johnson, 2002; Chapell, et al., 2005; Rana & Mahmood, 2010). Although emotionality in taking an examination is theoretically expected to be detrimental to academic achievement, some studies suggests that it may actually do the opposite. This current finding seems to concur with the results of Cassady and Johnson (2002), which indicated that physiological arousal can be beneficial to students. These physiological manifestations of anxiety may have been effective in arousing the participants to perform well in the licensure examination. It can be argued that high levels of emotionality can enhance the performance since it excites the body system above normal functioning capacity causing the individual to perform well (Coon & Mitterer, 2012).

Likewise, it should be noted that the current finding is under the condition for which the participants are assumed to have the same level of worry. This means that, should everyone have the same degree of worry, those who experience more physiological arousal are actually more likely to pass the licensure exam. While emotionality may emanate from worry itself (Deffenbacher, 1977), perhaps the kind of arousal which is not a product or side effect of worrying can actually be beneficial.

Furthermore, given the fact the participants were trained in allied health, they are probably knowledgeable of the physiological mechanisms of anxiety. Hence, they likely perceived emotionality as a result of normal bodily processes experienced by people, in response to challenges or threatening situations. They may have even capitalized on such arousing experience. Future researchers are encouraged to further investigate the inconsistencies in the effects of emotionality as a component of test anxiety. It will be good to be able to test

hypotheses about possible conditions when emotionality can be detrimental or beneficial to students.

With regard to the result of the moderation hypotheses, defensive pessimism somehow weakened the negative effects of worry on the licensure examination performance. This means that while being preoccupied with the negative outcomes of an exam can be detrimental to the actual performance, the strength of this relationship is weaker when individuals manage their worries and anxieties by preparing for the worst case scenarios (Norem & Cantor, 1986). This finding provides some evidence to the possible benefits students may get out of defensive pessimism (Norem & Cantor, 1986; Terada & Ura, 2015). Students who relatively set low expectations for themselves are less likely to experience the debilitating effect of anxiety on performance because they tend to prepare more (Norem & Cantor, 1986). Defensive pessimists are more deliberate in answering (as opposed to guessing), especially when they think of potential negative outcomes (Seery, West, Weisbuch, & Blascovich, 2008). Findings in the study of Riveiro (2014) also suggest that defensive pessimism is associated with the use of other desirable learning strategies.

Nevertheless, it should be noted that while defensive pessimism do buffer the negative effect of worry, results show that test anxiety still imposes its detrimental effect even at high levels of defensive pessimism, though to a lesser degree. This means that despite defensive pessimists' effort to use anxiety as motivation, such may not be enough to fully negate the detrimental effects to test anxiety. Despite being associated to many positive attributes (Riveiro, 2014), defensive pessimism is also related to negative student attributes. For example defensive pessimism has been associated with having performance-avoidance achievements goals (del Mar Ferradás, Freire, Núñez, Piñeiro, & Rosário, 2017). Further research is needed to confirm the consistency of defensive pessimism, as a protective strategy against test anxiety; and the conditions for which such role might hold true.

Based on the findings, some practical implications can be drawn. Considering that worry can be detrimental to test performance, measures to minimize worrying while preparing for or during the exam should be put into place. Teachers should create learning environments that promote confidence among students. Students who are preparing for an important examination should also be educated about the debilitating effects of worry on test performance. Furthermore, since worrying might be difficult to control, students should at least be taught how to utilize such the way defensive pessimists do, which is to use worry as motivation to prepare more for

upcoming examinations. Based on the current evidence, this should somehow mitigate the negative effects of worry on test performance. Since emotionality associated with taking tests was observed to have a positive influence on examination performance based on the current investigation, students should learn not impose negative interpretation over the physiological arousal they experience. Instead, they should think of capitalizing on such arousal as it may simply mean that their body is mustering various physiological resources in preparation for an important event, such as taking a high-stakes test.

References

Akca, F. (2011). The relationship between test anxiety and learned helplessness. *Social Behaviour and Personality, 31*(1), 101-112.

Arguelles, L., McCraty, R., & Rees, R. (2003). The heart in holistic education. *Encounter: Education for Meaning and Social Justice, 16*(3), 13-21.

Basavanna, M. (2000). *Dictionary of psychology.* New Delhi: Allied Publishers Ltd.

Berk, R., & Nanda, J. (2006). A ramdomized trial of humor effects on test anxiety and test performance. *Humor: International Journal of Humor Research, 19*(4), 425–454.

Cassady, J., & Johnson, R. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270-295.

Chapell, M., Blanding, Z., Silverstein, M., Takahashi, M., Newman,, B., Gubi, A., et al. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268-274.

Cole, J., & Osterlind, S. (2008). Investigating differences between low-and high-stakes test performance on a general education exam. *Journal of General Education, 57*, 119-130.

Coon, D., & Mitterer, J. (2012). *Introduction to psychology: Gateways to Mind and Behavior, 13e.* Belmont CA: Wadsworth.

Costello, C. (1976). *Anxiety and depression: The adaptive emotions.* Montreal: McGall-Queen's University Press.

Damer, D., & Melendres, L. (2011). "Tackling Test Anxiety": A group for college students. *The Journal for Specialists in Group Work, 36*(3), 163-177.

Deffenbacher, J. (1977). Relationship of worry and emotionality to performance on the Miller Analogies Test. *Journal of Educational Psychology, 61*(2), 191-195.

del Mar Ferradás, M., Freire, C., Núñez, J., Piñeiro, I., & Rosário, P. (2017). Motivational profiles in university students. Its relationship with self-handicapping and defensive pessimism strategies. *Learning and Individual Differences, 56*, 128-135.

Eysenck, M., & Calvo, M. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion, 6*(6), 409-434.

Garcia, T. (1995). The role of motivational strategies in self-regulated learning. *New Directions for Teaching and Learning, 1995*(63), 29-42.

Gray, J. (1991). Fear, panic, and anxiety: What's in a name? *Psychological Inquiry, 2*, 77–78.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Hong, E., & Karstensson, L. (2002). Antecedents of state test anxiety. *Contemporary Educational Psychology, 27*(2), 348-367.

Lerner, J., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology, 81*(1), 146-159.

Liebert, R., & Morris, L. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20*(3), 975-978.

Liepmann, D., Marggraf, C., Felfe, J., & Hosemann, A. (1992). Anxiety, action orientation, subjective state and situational aspects: A study of tank-lorry drivers. In K. Hagtvet, & T. Johnsen, *Advances in Test Anxiety Research* (Vol. 7, pp. 130-141). Amsterdamllisse: Swets and Zeitlinger.

Liew, J., Lench, H., Kao, G., Yeh, Y., & Kwok, O. (2014). Avoidance temperament and social-evaluative threat in college students' math performance: A mediation model of math and test anxiety. *Anxiety, Stress, & Coping, 27*(6), 650-661.

McCraty, R. (2005). Enhancing emotional, social, and academic learning with heart rhythm coherence feedback. *Biofeedback, 33*(4), 130-134.

McDonald, A. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology, 21*(1), 89-101.

Morris, L., & Liebert, R. (1970). Relationship of cognitive and emotional components of test anxiety to physiological arousal and academic performance. *Journal of Consulting and Clinical Psychology, 35*(3), 332-337.

Morris, L., Smith, L., Andrews, E., & Morris, N. (1975). The relationship of emotionality and worry components of anxiety to motor skills performance. *Journal of Motor Behavior, 7*(2), 121-130.

Moser, J., Moran, T., Schroder, H., Donnellan, B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: A meta-analysis and conceptual framework. *Frontiers in Human Neuroscience, 7*, 1-19.

Norem, J. (2001). Defensive pessimism, optimism, and pessimism. In E. Chang, *Optimism and Pessimism: Implications for Theory, Research, and Practice* (pp. 77-100). Washington, DC: American Psychological Association.

Norem, J., & Cantor, N. (1986). Defensive pessimism: Harnessing anxiety as motivation. *Journal of Personality and Social Psychology, 51*(6), 1208-1217.

Pintrich, P. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385-407.

Rana, R., & Mahmood, N. (2010). The relationship between test anxiety and academic achievement. *Bulletin of Education and Research, 32*, 63-74.

Riveiro, J. (2014). Optimistic and defensive-pessimist students: differences in their academic motivation and learning strategies. *The Spanish Journal of Psychology, 17*, 1-18.

Schwarzer, R. (1984). Worry and emotionality as separate components in test anxiety. *Applied Psychology, 33*(2), 205-220.

Seery, M., West, T., Weisbuch, M., & Blascovich, J. (2008). The effects of negative reflection for defensive pessimists: Dissipation or harnessing of threat? *Personality and Individual Differences, 45*(6), 515-520.

Spielberger, C. (1980). *Test Attitude Inventory: Preliminary Professional Manual.* Palo Alto, CA: Consulting Psychologists Press.

Terada, M., & Ura, M. (2015). Positive thinking impairs subsequent self-regulation: Focusing on defensive pessimists and optimists. *Journal of Educational and Developmental Psychology, 5*(2), 28-38.

Thompson, T., & le Fevre, C. (1999). Implications of manipulating anticipatory attributions on the strategy use of defensive pessimists and strategic optimists. *Personality and Individual Differences, 26*(5), 887-904.

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2), 92-104.

Wong, S. (2008). The relations of cognitive triad, dysfunctional attitudes, automatic thoughts, and irrational beliefs with test anxiety. *Current Psychology, 27*, 177-191.

Zeidner, M. (1998). *Test Anxiety: The State of the Art.* New York: Plenum Press.

Educational Measurement and Evaluation Review (2018), Vol. 9 Issue 1

Assessing the Construct Validity of the Locus-of-Hope Scale

Dominique T. Rivera
Philippine Normal University

Leny G. Gadiana
University of Sto. Tomas

Abstract

The Locus-of-Hope Scale (Bernardo, 2010) was developed as a measure of the locus-of-hope constructs (internal locus, external locus-parent, external locus-peer, external locus-spiritual). This study aimed to examine the construct validity of the Locus-of-Hope Scale using alternative approaches in establishing convergent and discriminant validity. A sample of 1, 214 Filipino university students participated in the study and their responses on the Locus-of-Hope Scale were analysed using confirmatory factor analysis. Results indicated that the correlated four-factor structure of the Locus-of-Hope Scale is valid based on acceptable fit indices. Results also indicated strong support for the discriminant validity of the Locus-of-Hope Scale, but three of the four subscales were found to have convergent validity issues related to their average variance extracted (AVE). Implications of the findings and recommendations for future research are discussed.

Keywords: locus-of-hope, Locus-of-Hope Scale, construct validity, convergent validity, discriminant validity

Introduction

Snyder's hope theory (1994; 2000) describes hope as a cognitive motivational system that allows a person to engage in goal-directed behaviour even when facing impediments. The research literature on hope indicates that hope is associated with a number of adaptive psychological outcomes. For

instance, it has been reported that hopeful people are more likely to have more positive adaptation to stress (Ong, Edwards, & Bergeman, 2006), stronger well-being (Shorey, Little, Snyder, Kluck, & Robitschek, 2007), and more positive affect (Steffen & Smith, 2013). In the educational domain, hopeful persons are more likely to have better academic performance (Rand, Martin, & Shea, 2011; Snyder et al., 2002). It has also been reported that hope-based interventions are effective in promoting positive psychological outcomes (Cheavens, Feldman, Gum, Michael, & Synder, 2006; Feldman & Dreher, 2012). These findings provide strong evidence that hopeful thinking is a desirable disposition for people to have. Therefore, the assessment of hope and the use of self-report instruments to facilitate assessment are essential.

The Locus-of-Hope Scale (Bernardo, 2010) is a recently developed self-report measure of hope that is based on the locus-of-hope model (Bernardo, 2010; Bernardo, 2014) which conceptualizes trait hope as having an internal or external locus. Bernardo (2010) extended Snyders's hope theory by proposing that hope may also be grounded on persons or agents outside of the individual. Bernardo (2014) explained that the notion of external loci of hope is consistent with the argument that a conjoint model of agency may exist in collectivist cultures that highlight the roles of other people in a person's goal attainment. Bernardo (2010) elaborated that in the internal locus-of-hope, the agent of goal-attainment cognitions is the individual, whereas in the external locus-of-hope, the agents of goal-attainment cognitions are significant people or forces external to the individual. Bernardo (2010) further proposed that the external locus-of-hope has three sub-dimensions: external locus-family (hope placed on one's family), external locus-peer (hope placed on peers or friends), and external locus-spiritual (hope placed on God or a superior spiritual being). Bernardo (2010; 2016) also argued that external loci- of- hope may even be more important for people in collectivist cultures. Thus, the LHS was developed with four subscales corresponding to the four locus-of-hope constructs (Bernardo, 2010). The validity of the Locus-of-Hope Scale was first examined through confirmatory factor analyses (CFA) of data from Filipino university students in the Philippines, wherein results supported a four-factor structure consistent with the proposed locus-of-hope dimensions (Bernardo, 2010). In the same study, the four-factor structure was also supported by the differential relations of the internal and external loci-of-hope on individual-level individualism and collectivism. The four-factor structure of the Locus-of-Hope Scale was further validated among young Filipino adolescents (Bernardo, 2014), and the results also supported the four-factor structure. In the same study, the Locus-of-Hope

Scale was also found to have measurement invariance across sex. The four-factor structure was also confirmed in the Chinese version (Du, Bernardo, &Yeung, 2015) and in the short-form and Filipino version (Bernardo & Estrellado, 2014) of the Locus-of-Hope Scale.

Construct validity pertains to the extent to which a set of measured indicators or items truly represent the theoretical latent construct those indicators are supposed to measure and has four components: convergent validity, discriminant validity, nomological validity, and face validity (Hair, Black, Babin, & Anderson, 2010). While a number of studies reported acceptable psychometric properties for the Locus-of-Hope Scale (e.g. Bernardo, 2010; 2014; Gadiana & David, 2015), the scale can still benefit from additional psychometric analysis of construct validity. First, the reliability for some of the subscales seems to be inadequate as indicated by marginal Cronbach's alpha values. For example, Bernardo (2014) reported that the internal locus subscale has an alpha value of .62 and Du and King (2013) reported that external locus-peer subscale has an alpha value of .71. Moreover, the use of Cronbach's alpha to determine reliability has been criticized because it can underestimate reliability (Sijtsma 2009) and may not be compatible with multi-dimensional scales (Teo & Fan, 2013). It is important that the reliability of the subscales be examined using alternative measures like composite reliability which is a better measure of reliability compared to Cronbach's alpha (Wong & Lo, 2012). Second, none of the studies that used the Locus-of-Hope Scale examined the average variance extracted (AVE) of the subscales. The AVE represents "a summary measure of convergence of the set of variables as a whole that represents a latent construct" (Wong & Lo, 2012, p. 403). Since the AVE is a more conservative indicator of validity (Teo & Jarupunphol, 2015), it is also important that the validity of the Locus-of-Hope Scale be assessed using AVE.

The present study reports the results of an assessment of the construct validity of the Locus-of-Hope Scale (LHS) using a Philippine sample. The goal of the study is to investigate the construct validity of the LHS through the use of more conservative approaches like using composite reliability and average variance extracted (AVE) values in order to provide stronger evidence for the convergent and discriminant validity of the LHS. The focus on convergent and discriminant validity is grounded on the argument that convergent and divergent validity are the most essential components of construct validity (Wong & Lo, 2012). Establishing reliability through composite reliability coefficients may also provide a more accurate picture of the reliability of the LHS.

Method

Measure

The Locus-of-Hope Scale (LHS). The original English version of the Locus-of-Hope Scale (Bernardo, 2010) was used in the study. Each of the locus-of-hope construct is measured by a corresponding subscale with eight (8) items. Utilizing a 4-point Likert-type scale, the LHS requires respondents to indicate the extent to which an item describes them using a scale from 1 (*definitely false*) to 4 (*definitely true*). Aside from the 32 locus-of-hope items, the LHS contains eight filler items. The following are sample items: "I can think of many ways to get the things in life that are important to me" (internal locus), "My parents have lots of ways of helping me attain my goals." (external locus-parent),"With the help of my friends, I am confident that I can reach my goals in life" (external locus-peer), and "God has many different ways of letting me attain my goals" (external locus-spiritual). All items are positively stated and stronger agreement with an item indicates higher level on the locus-of-hope construct that the item represents. Subscale scores were obtained by computing the mean scores of the participants' responses across the items in each subscale.

Participants

The present study used a convenience sample of 1, 214 undergraduate students from a university in the National Capital Region (NCR) of the Philippines. The language of instruction used in the aforementioned university is English. Since English is the medium of instruction in Philippine schools from secondary education to college, it was assumed that the participants of the study can read and undertand the items of the LHS. Furthermore, the initial validation of the English version of the LHS was also done with Filipino university students (Bernardo, 2010). There were 819 (67.46%) female participants and 395 (32.54%) male participants. The participants' ages ranged from 16 to 22 years (M = 18.69 years; SD = 0.90). In terms of religion, 963 (79.32%) of the participants reported that they are Catholic. The participants came from various educational majors and participating classes were selected in coordination with academic offices and faculty. All participants responded to the LHS during class hours. The participants were informed of the purpose of the study and informed consent was sought from the participants prior to administration of the LHS.

Data Analysis

A series of statistical analysis was performed on the data of the participants' responses on the LHS. First, descriptive statistics of the items were computed using SPSS 20. Second, the factorial structure of the LHS was examined as a measurement model through CFA. In the CFA, the covariance matrix of the data was analyzed through Maximum Likelihood Estimation (MLE) using the software AMOS 20. The four-factor structure of the LHS was assessed by determining whether an item loaded on its hypothesized latent factor, and whether the correlated four-factor structure of the locus-of-hope dimensions obtained a good fit with the data. To evaluate the fit of the model tested, a number of goodness-of-fit indices were considered: Chi square ($\chi 2$), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and root mean square error of approximation (RMSEA). Model fit was evaluated using the following criteria: the $\chi 2$ should not be significant, CFI and TLI should at least be .90 and RMSEA should not be higher than .08 (David, 2012). Convergent validity was assessed by examining the item factor loadings as indicated by standardized parameter estimates (SE), construct reliability as measured by composite reliability coefficient (CR), and the average variance extracted (AVE) of each of the locus-of-hope subscales. Discriminant validity was assessed by comparing the square root of the AVE for a construct with all the bivariate correlations of that construct with each of the other constructs. Moreover, the maximum shared variance (MSV) and average shared variance (AVS) of each construct was also computed. The CR, AVE, MSV, and AVS values were computed from CFA outputs using Microsoft Excel 2010.

Results

Table 1
Descriptive statistics for the items of the LHS subscales

Items	M	SD	Skewness	Kurtosis
Internal Locus				
Item 1	3.13	.59	-.21	.47
Item 6	3.22	.64	-.32	-.23
Item 14	3.19	.62	-.21	-.28

Item 20	3.32	.58	-.28	-.22
Item 23	3.20	.64	-.27	.29
Item 27	3.23	.58	-.16	-.04
Item 30	3.30	.64	-.49	-.15
Item 40	3.09	.63	-.31	.37
External Locus-Parent				
Item 3	3.60	.56	-1.02	.04
Item 7	3.38	.64	-.58	-.33
Item 11	3.40	.64	-.63	-.41
Item 16	3.33	.64	-.45	-.61
Item 21	3.37	.60	-.46	-.28
Item 24	3.31	.60	-.26	-.63
Item 32	3.35	.60	-.36	-.54
Item 39	3.26	.64	-.31	-.61
External Locus-Peer				
Item 5	2.76	.78	-.23	-.32
Item 10	3.25	.61	-.29	-.15
Item 13	3.13	.65	-.35	.19
Item 19	3.03	.62	-.19	.24
Item 26	2.91	.63	-.29	.43
Item 33	3.04	.63	-.27	.36
Item 35	3.04	.63	-.33	.57
Item 38	2.94	.65	-.18	.01
External Locus-Spiritual				
Item 2	3.68	.53	-1.38	.95
Item 9	3.64	.54	-1.13	.27
Item 15	3.72	.50	-1.50	1.29

Item 17	3.58	.58	-1.03	.07
Item 22	3.67	.52	-1.30	.70
Item 28	3.65	.52	-1.07	.02
Item 34	3.64	.53	-1.06	.06
Item 36	3.62	.55	-1.09	.17

Descriptive Statistics

The descriptive statistics of the items in the four subscales of the LHS are presented in Table 1. The mean scores of all 32 items were above the midpoint value of 2.5, indicating that the participants positively endorse the items. The standard deviation scores indicate a narrow spread of the scores around the mean. Based on Kline's (2005) recommendations that skewness and kurtosis values should be within | 3 |and | 10 | respectively, the data were assumed to have univariate normality. Since the use of the MLE approach in CFA requires data to have multivariate normality, the Mardia's normalized multivariate kurtosis value was also examined. Following the approach applied by Teo and Noyes (2014), the data was assumed to have multivariate normality since the obtained Mardia's coefficient of 176.08 is lower than the value of 1, 088 computed from the formula $p\ (p + 2)$ where p equals the number of observed variables

Factor Structure

The obtained Hoelter's (1983) Critical *N* for the data is 391 (.01) which is lower than the total number of participants in the study. This suggests that the sample size used in the study is sufficiently large for testing the measurement model. The CFA of the correlated four-factor structure of the LHS showed that all items loaded significantly on their hypothesized latent factors and yielded the following fit indices: χ^2 [(458, N= 1,214) = 1, 649, p< 0.001]; CFI = .93, TLI = .93, RMSEA= .046 (CI: .044; .049), SRMR = .050. All fit indices met the criteria for a good fitting model, except for the χ^2 value. Since a significant χ^2 value is expected for a model with a sample size greater than 250 and at least 30 observed variables (Hair et al., 2010), it can be said that the correlated four-factor structure achieved a relatively good fit with the data.

Table 2
Item standardized estimates (SE), composite reliability (CR), and AVE of the LHS

Subscale Items	SE	CR	AVE
Internal Locus		.79	.32
Item 1	.51		
Item 6	.58		
Item 14	.56		
Item 20	.61		
Item 23	.55		
Item 27	.61		
Item 30	.48		
Item 40	.60		
External Locus-Parent		.88	.49
Item 3	.61		
Item 7	.73		
Item 11	.72		
Item 16	.73		
Item 21	.60		
Item 24	.74		
Item 32	.72		
Item 39	.71		
External Locus-Peer		.88	.48
Item 5	.52		
Item 10	.58		
Item 13	.72		
Item 19	.67		
Item 26	.76		
Item 33	.74		
Item 35	.71		
Item 38	.76		
External Locus-Spiritual		.92	.60
Item 2	.76		
Item 9	.77		
Item 15	.83		
Item 17	.72		
Item 22	.78		

Item 28	.80
Item 34	.75
Item 36	.79

Convergent and Discriminant Validity

There are three ways to examine the convergent validity of a measure (Hair et al., 2010). First is the size of the *factor loadings*, wherein high factor loadings indicate convergence of the items on the latent construct that they are supposed to measure. Second is *reliability*, wherein the internal consistency of a set of items indicates convergence of those items on the latent construct. Third is the AVE which measures the amount of variance captured by a construct in relation to the amount of variance attributed to measurement error (Teo & Jarupunphol, 2015). Convergent validity is deemed adequate if item factor loadings as indicated by standardized parameter estimates is at least 0.50, the composite reliability is at least .70, and the AVE is at least .50. Table 2 shows the results of the three measures of convergent validity used in this study. The results indicate that the parameter estimates for all items are at 0.50 or higher, except for one item in the internal locus construct. This means that the LHS has adequate convergent validity at the item level. The composite reliability coefficients were all above .70, indicative of good reliability and convergent validity at the construct level. However, the AVE values were less than satisfactory as only the AVE of the external locus-spiritual subscale met the cut-off score of 0.50. While the AVE values of the external locus-parent and external locus-peer subscales were also unsatisfactory as they were below the recommended guideline, the AVE of the internal locus subscale was very low. To explore the low AVE of the internal locus construct, the R^2 values of all the items in this construct were obtained and results showed that the percent of variance explained ranges from .26 to .37. This is problematic as no single item has an R^2 value of at least .50. In contrast, some items in the external locus-family and external locus-peer have R^2 values of at least .50. Nevertheless, the acceptable values of the item parameter estimates and construct reliability coefficients of the internal locus, external locus-family, and external locus-peer subscales suggest sufficient convergent validity. This means that the four-factor structure of the LHS has acceptable convergent validity, but with weaker AVE values.

The discriminant validity of a measure can be assessed by examining the

correlation between its latent constructs (Hair et al., 2010). A very high correlation between constructs suggests lack of discriminant validity, while low to moderate correlation is indicative of good discriminant validity. Inter-scale correlations of the four subscales yielded correlation scores ranging from .27 to .52, indicating sufficient discriminant validity. This serves as evidence that the four locus-of-hope constructs are conceptually unique and distinct dimensions of locus-of-hope. An alternative approach to assessing discriminant validity is comparing the square root of the AVE for a construct with all the bivariate correlations of that construct with all the other constructs (Teo & Noyes, 2014). If the square root of the AVE of the construct is higher than all of the bivariate correlations of that construct, then the construct is deemed to have discriminant validity. Table 3 presents the square roots of the AVE of the locus-of-hope subscales in parenthesis. The square root of the AVE for a construct was compared with the correlations of that construct with each of the other three locus-of-hope constructs. Results indicated that all constructs appear to have satisfactory discriminant validity. There is also discriminant validity if the AVE of a subscale is higher than the computed maximum shared variance (MSV) and average shared variance (ASV) of all subscales. The results yielded MSV and AVS scores that were lower than the respective AVE of each subscale. In summary, it appears that the LHS has strong discriminant validity.

Table 3
Inter-scale correlations, square roots of the AVEs, maximum shared variance (MSV), and average shared variance (ASV) of the LHS

Subscale	1	2	3	4	AVE	MSV	ASV
1 Internal Locus	(.56)				.32	.27	.21
2 External Locus-Parent	.52	(.70)			.49	.27	.24
3 External Locus-Peer	.51	.50	(.69)		.48	.26	.19
4 External Locus-Spiritual	.32	.45	.27	(.77)	.60	.20	.13

Discussion

The present study performed psychometric analysis on the Locus-of-Hope Scale in order to obtain additional evidence on its construct validity. In terms of factorial structure, results are consistent with previous research documenting the structural validity of the four-factor structure of the English version of the LHS (e.g. Bernardo, 2010; 2014). This indicates that the four locus-of-hope constructs can be differentiated from the students' responses to the LHS. More importantly, the results of the study provide strong support for the discriminant validity of LHS. Interestingly, evidence for the separation and distinctiveness of the locus-of-hope constructs was strengthened by using the alternative approach of analyzing the AVE of a measure.

The analysis of the convergent validity of the LHS got mixed results. Adequate convergent validity is evident based on obtained item standardized estimates where only one of the 32 locus-of-hope items did not meet the minimum criteria for acceptable factor loading. Convergent validity is also evident in the obtained composite reliability coefficients. The composite reliability coefficients for the LHS were all satisfactory and exceeded the reliability estimates based on Cronbach's alpha that were reported in previous studies (Bernardo, 2014; Bernardo et al., 2015). Taken together, the factor loadings and reliability estimates indicate convergence of items on the locus-of-hope construct that the items are supposed to measure. However, the LHS demonstrated low AVE for three locus-of-hope subscales. Low AVE is a validity problem because it indicates that measurement errors explain more variance in the items than the latent construct to which the items are loaded. As the study is the first to examine the AVE of the LHS subscales, this validity problem of the LHS items and subscales was not detected in previous studies that relied primarily on factor loadings and model fit to validate the LHS. The convergent validity problem observed on some of the subscales of the LHS seems to suggest the need to review the items of these subscales to determine if item revision or item construction would be necessary to improve the scale and obtain stronger convergent validity. Another plausible strategy is to develop a short-form of the LHS by removing problematic items and retaining items that contribute strongly to convergent validity. Previously, a short form of the Filipino version of LOH was developed (Bernardo & Estrellado, 2014).

In general, the construct validation approach used in this study provides adequate support for the construct validity of the LHS and its usefulness as a measure of hope. The LHS is particularly useful for researchers who are

interested in determining the role of internal and external loci of hope in the educational and psychological experiences of students. The LHS may also be a useful tool for school counsellors in assessing the hopeful cognitions of university students which can provide inputs to school counselling and career development programs. Nevertheless, more research on the psychometric properties of the LHS is warranted especially on how the convergence of the items in each subscale can be improved. Cross-cultural validation of the English version of the LHS is also needed and it is imperative to determine its validity in more individualist cultures in order to expand its utility as a measure of dispositional hope.

A major limitation of this study was that only Filipino students from one university served as respondents. Future research could sample a wider range and more diverse group of respondents for stronger generalization of the findings. Moreover, no information was sought on whether the participants have current or prior psychopathology. Thus, the results do not offer support for the utility of the LHS in clinical samples. The utility of the LHS on assessing the hopeful cognitions of clinical samples must be explored in future studies. The present study also did not take into account the nomological validity of the Locus-of-Hope Scale. While the nomological validity of the Locus-of-Hope Scale can be inferred from the results of studies that show the differential relations of the locus-of-hope constructs with adaptive outcomes like life satisfaction (Du et al., 2015; Du & King, 2013), use of learning strategies (Bernardo, Salanga, Khan, & Yeung, 2015), and future goals (Gavilano, Nalipay, & David, 2018), there is still a need to investigate how the locus-of-hope constructs relate with other psychological constructs. An important line of inquiry is determining the association of the locus-of-hope constructs with students' academic outcomes While the association of academic achievement and hope drawn from within oneself has been examined (e.g. Rand et al., 2011; Synder et al., 2002), there is a dearth of studies on how the three external locus-of-hope dimensions relate to academic success. One exemption is the study of Lucas and Ouano (2018) who examined the predictive influence of the locus-of-hope dimensions on the academic achievement among Filipino college indigent students.

In spite of the aforementioned limitations, the present study contributes to the literature on locus-of-hope by providing additional evidence for the construct validity of the LOH. In future research involving the use of self-report instruments like the LHS, the construct validity of the measurement model should be established beyond factor loadings and fit indices. This can be done

by obtaining evidence of convergent and discriminant validity, which are considered as the most essential components of construct validity (Wong & Lo, 2012). By establishing the construct validity of an assessment tool or instrument, one can be more certain that the instrument is really measuring what it is supposed to measure.

References

Batara, J. B. (2015). Overlap of religiosity and spirituality among Filipinos and its implications towards religious prosociality. *International Journal of Research Studies in Psychology, 4*(3), 3-21.

Bernardo, A. B. (2010). Extending hope theory: Internal and external locus of trait hope. *Personality and Individual Differences*, *49*, 944–949.

Bernardo, A. B. I. (2010). Exploring Filipino adolescents' perceptions of the legitimacy of parental authority over academic behaviors. *Journal of Applied Developmental Psychology*, *31*, 271–280.

Bernardo, A. B. (2014). Hope in early adolescence: Measuring internal and external locus-of-hope. *Child Indicators Research*, *8 (3),* 699-715.

Bernardo, A. B., & Estrellado, A. F. (2014). Measuring hope in the Philippines: Validating the short version of the Locus-of-Hope Scale in Filipino. *Social Indicators Research*, *119* (3), 1649-1661.

Bernardo, A. B., Salanga, M.G., Khan, A., & Yeung, S. (2015). Internal and external loci-of-hope predict use of individual and collaborative learning strategies: Evidence from university students in four Asian cities. *The Asia-Pacific Education Researcher.* doi:10.1007/s40299-015-0249-y.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp.136-162). Beverly Hills, CA: Sage.

Curry, L., Snyder, C. R., Cook, D., Ruby, B., & Rehm, M. (1997). Role of hope in academic and sports achievement, *Journal of Personality and Social Psychology, 73* (6), 1257-1267.

Cheavens, J. S., Feldman, D. B., Gum, A., Michael, S. T., & Snyder, C. R. (2006). Hope therapy in a community sample: A pilot investigation. *Social Indicators Research, 77(1),* 61–78.

David, A.P. (2012). Structural validation of the 3 x 2 achievement goal model. *Educational Measurement and Evaluation Review*, *3*, 50-59.

Day, L., Hanson, K., Maltby, J., Proctor, C., & Wood, A. (2010). Hope uniquely

predicts objective academic achievement above intelligence, personality, and previous academic achievement. *Journal of Research in Personality, 44,* 550-553.

Drach-Zahavy, A., & Somech, A. (2002). Coping with health problems: The distinctive relationships of hope sub-scales with constructive thinking and resource allocation. *Personality and Individual Differences, 33,* 103-117.

Du, H., Bernardo, A. B. I., & Yeung, S. (2015). Locus-of-hope and life satisfaction: The mediating roles of personal self-esteem and relational self-esteem. *Personality and Individual Differences*, *83*, 228–233.

Du, H., & King, R. B. (2013). Placing hope in self and others: Exploring the relationships among self-construals, locus of hope, and adjustment. *Personality and Individual Differences*, *54*, 332–337.

Feldman, D. B., & Dreher, D. E. (2012). Can hope be changed in 90 minutes? Testing the efficacy of a single-session goal-pursuit intervention for college students. *Journal of Happiness Studies, 13*, 745–759.

Gadiana, L.**,** & David, A.P. (2015). Rasch analysis of the Locus-of-Hope Scale. *Educational Measurement and Evaluation Review, 6,* 32-37.

Gavilano, V., Nalipay, M., & David, A.P. The role of hope in promoting society-oriented future goal. *The Normal Lights*, *12* (1), 185-198.

Gerbing, D., & Anderson, J. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research, 11*, 572-580.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall International.

Hoe, S.L. (2008). Issues and procedures in adopting structural equation modelling technique. *Journal of Applied Quantitative Methods, 3,* 76-83.

Hoelter, J. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11,* 325-344.

King, R., & Ganotice, F. (2014). The social underpinnings of motivation and achievement: Investigating the role of parents, teachers, and peers on academic outcomes. *The Asia-Pacific Education Researcher, 23* (3), 745-756.

King, R., & Ganotice, F., & Watkins, D. (2014). A cross-cultural analysis of achievement and social goals among Chinese and Filipino students. *Social psychology of education,17,* 439-455.

Kline, R.B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Lucas, R.I., & Ouano, J. (2018). Hope and academic achievement among young Filipino college indigent students. *Asia-Pacific Social Science Review, 17* (3), 1-14.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods,7,* 64–82.

Ong, A., Edwards, L., & Bergeman, C.S. (2006). Hope as a source of resilience in later adulthood. *Personality and Individual Differences*, *41*, 1263-1273.

Rand, K., Martin, A., & Shea, A. A. (2011). Hope, but not optimism, predicts academic performance of law students beyond previous academic achievement. *Journal of Research in Personality, 45,* 683-686.

Shorey, H., Little, T., Snyder, C. R., Kluck, B., & Robitschek, C. (2007). Hope and personal growth initiative: A comparison of positive, future-oriented constructs. *Personality and Individual Differences*, *43*, 1917-1926.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's α. *Psychometrika, 74,*107-120.

Snyder, C. R. (1994). *The psychology of hope.* New York: Free Press.

Snyder, C. R. (1995). Conceptualizing, measuring, and nurturing hope. *Journal of Counseling and Development, 73,* 355–360.

Snyder, C. R. (Ed.). (2000). *Handbook of hope: Theory, measures, and applications.* San Diego, CA: Academic Press.

Snyder, C. R., Harris, C., Anderson, J. R., Holleran, S. A., Irving, L. M., Sigmon, S. T., et al. (1991). The will and the ways: Development and validation of an individual differences measure of hope. *Journal of Personality and Social Psychology*, *60*, 570–585.

Snyder, C. R., Shorey, H. S., Cheavens, J., Pulvers, K. M., Adams, V. H., & Wiklund, C.(2002). Hope and achievement success in college. *Journal of Educational Psychology*, *94,* 820–826.

Steffen, L., & Smith, B. (2013). The influence of between and within-person hope among emergency responders on daily affect in a stress and coping model. *Journal of Research in Personality, 47,* 738-747.

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher, 22* (2), 209-213.

Teo, T., & Jarupunphol, P. (2015). Dhammic technology acceptance model: Extending the TAM using a condition of attachment in Buddhism. *Journal of Educational Computing Research, 52* (1), 136-151.

Teo, T., & Noyes, J. (2014). Explaining the intention to use technology among pre-service teachers: A multi-group analysis of the unified theory of acceptance and use of technology. *Interactive Learning Environments, 22 (1),* 51-66.

Wong, A. K., & Lo, E. S. (2012). Assessing the construct validity of the

Conceptions for Teaching and Learning Questionnaire (CTLQ) for Chinese university students in Hong Kong: Going beyond the use of goodness of fit indices. *The Asia-Pacific Education Researcher, 21(2),* 402-413.