ISSN 20194-5876

Educational Measurement and Evaluation Review

VOLUME 7 ISSUE 2
DECEMBER 2016

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
◈ December 2016, Vol. 7 Issue 2

Educational Measurement and Evaluation Review (2016), Vol. 7 Issue 2

Academic Delay of Gratification, Academic Achievement, and Need for Affiliation of Selected High School Students

Ryan Francis O. Cayubit
Christine Allen D. Cadacio
Mary Pauline Therese O. Chua
Van Alistair H. Faeldon
WilletteValjean P. Go
Marc Kristoffer C. Verdan
University of Santo Tomas, Manila, Philippines

Abstract
The study looked into the ability of academic delay of gratification (e. g. intentionally miss out a social event such as parties and hanging out in order to be able to focus on their studies) and need for affiliation (e. g. establishing and managing close interpersonal relationships with others) to predict the academic achievement (e. g. average grade of all subjects during the first grading period of the academic year) of high school students. A sample of 1,021 Filipino fourth year high schools students from selected private and public high schools in Metro Manila participated in this study. Results showed that academic achievement was positively predicted by academic delay of gratification but negatively predicted by need for affiliation an indication of the ability of high school students to prioritize goals.

Keywords: academic delay of gratification, need for affiliation, academic achievement

Introduction

In the world of the academe, the everyday life of students is not an easy one. They are often faced with

academic challenges where teachers require them to turn in large amount of schoolwork and at the same time expect them to perform at par within the established standards of their respective schools. Though most of their time is spent in school, one cannot deny the fact that students are also involved in other pursuits or activities that typically involve their friends and other social groups. This is where the challenge comes in, as students must learn to balance their schoolwork, time, and other demands by their friends and other social groups. However, reality dictates that this is easier said than done, as oftentimesstudents tend to fail at this task resulting in grave consequences in their academic and social lives. Thus the task of trying to maintain a healthy school and social life can result to strife resulting to the need to compromise. The compromise revolves around the perspective of choosing one activity over the other (school work or friends or social groups) based on the subjective perception of which one is considered more important or a priority that is often aligned with their goals. Just like the task of balancing demands, the choice that students need to make is not an easy one. Their choice is often influenced by upbringing, values, personality, and skills like self-regulation.

According to Zimmerman (1994), self-regulation is a process whereby students utilize cognitions, behaviors, and affects that are systematically oriented towards the achievement of goals. Thus it can be used to help students decide on what they would like to do as this process has the capacity to direct and control their actions and emotions (Eggen & Kauchak, 2012), as well as the underlying motives under these overt expressions (Bandura, 1986; Zimmerman, 1989). The application of this process appears wide and students do not only use it in terms of making choices but also to help them in their schoolwork since self-regulation enables them to dynamically and strategically act on their learning and not just passively receive knowledge (Pintrich & Schrauben, 1992). In a general way, students with self-regulated behavior are viewed to pursue long-term goals not only in learning but in other areas of their lives as well. Just like other psychological constructs, self-regulation is seen as

something that is multidimensional and one of its dimensions is what is popularly known as delay of gratification (Zimmerman, 2000; Eggen & Kauchak, 2012).

Delay of gratification refers to the voluntary turn down of immediate gratification and to endure self-imposed delays of reward involving self-control patterns of behavior (Mischel, 1996; Bembenutty & Karabenick, 1998). When applied to the school setting, delay of gratification becomes more specific and is referred to as academic delay of gratification. This involves the students' postponement of immediately available opportunities to satisfy impulses in favor of pursuing important academic rewards, goals, and intentions that are temporally remote but ostensibly more valuable (Bembenutty & Karabenick, 1998). This may mean students declining invites to social gatherings or events in favor of their academic work and students who are able to do this even when faced with high social pressure is said to possess high academic delay of gratification. To understand the construct better, Bembenutty and Karabenick (1998) stressed that academic delay of gratification must be viewed from three specific perspectives. According to them, academic delay of gratification is a positive ability or a sign of competence among students that can be developed over time because of the continuous use learning-delay relevant strategies like controlling one's attention or the use of verbal motivation as a reminder to study. The continuous use of academic delay of gratification among students can also help develop their ability for self-control and will power (Mischel, 1996).Academic delay of gratification is also tied to one's personality and is considered as a stable disposition (Bembenutty, 1999; Zhang, Karabenick, Maruno, & Lauermann, 2011) and its continuous use is also an indication that students are more oriented towards their future and the fulfillment of their goals (Mischel, 1981). Lastly, it can be used as a strategy to achieve long-term goals and are activated by motivational determinants (Bembenutty & Karabenick, 1998; Bembenutty, 2007). This was elaborated further by Arabzadeh, Kadivar, and Dlavar (2012) when they identified specific types of strategies involved when academic delay of gratification is used and that these

strategies work and act together when it is exercised (Pintrich & De Groot, 1990; Bembenutty & Karabenick, 2004). According to them, students make use of strategies that are cognitive (rehearsal, elaboration, organization, critical thinking) and metacognitive (planning, goal-setting, monitoring, self-evaluation) in nature. Resource-management strategies (time management, self-seeking) and motivational strategies (self-efficacy, self-satisfaction) may also be involved when academic delay of gratification is used (Arabzadeh et al., 2012).

One's ability to delay gratification in the academic setting could be affected by factors such as an individual's external environment and needs. Needs could affect the personality and behavioral dynamics since these serve as motivating factors of an individual. The most notable literature on this topic is McClelland's motivational needs. According to his theory, there are three types of motivational needs: need for achievement, need for power, and need for affiliation (Harrell & Stahl, 1983). For the present study, the researchers opted to focus on the students' need for affiliation. This particular need puts emphasis on interpersonal relationships since it motivates individuals to seek and establish warm interpersonal relationships with others within their immediate environment (Harrell & Stahl, 1983). The need for affiliation is tied to personality and those individuals with a high need for affiliation are often driven towards creating and managing close interpersonal relationships driven by their desire for social contact and belongingness with others because this allows them to feel secure and connected (Harrell & Stahl, 1983; Wiesenfeld, Raghuram, & Garud, 2001; Slabbinck, De Houwer, & Kenhove, 2012) and to experience acceptance, friendship, love (Hofer & Busch, 2011) and happiness (Schüler, Job, Fröhlich, & Brandstätter, 2008). Other rewards include experiencing gratification through social communion with others (Wiesenfeld, et al., 2001). This explains why students tend to value more their relationship, the views and opinions of their friends or peers over other people. This type of need is also subjective in nature because its intensity tends to differ from one person

to another (Wiesenfeld et al., 2001; Schüler, et al., 2008). It is also universal because it is felt or experienced by everybody and more often than not people direct or exert more effort in maintaining and enhancing their interpersonal relationships (Mathieu, 1990). When applied to the school setting, this is where the conflict arises particularly when the need for affiliation of students is high resulting to their tendency to prioritize this need over their other needs and goals that are academic in nature.

Both variables are deemed important since studies have shown that both academic delay of gratification and need for affiliation can affect student performance. This is a concern since most of the time, student performance has become a basis for promotion to the next level, access to scholarships, access to universities and as one of the bases for hiring and selection for jobs. Student performance may pertain to their academic achievement, which denotes their knowledge and learning gained over a period of time while they are in school (Russell & Airasian, 2012).Past researches show that academic delay of gratification is positively related to academic achievement (Bembenutty, 2011; Herndon, Bembenutty, & Gill, 2015) and that those skills incorporated within the use of academic delay of gratification like cognitive, metacognitive, and resource management strategies also predict academic achievement (Bembenutty & Karabenick, 1998). This is an indication that these students have successfully resisted temptations that are immediately gratifying in order to increase the likelihood of accomplishing temporally remote and presumably more important goals related to their academics (Bembenutty & Karabenick, 2004).

Studies have also documented the link between need for affiliation and academic performance. A negative relationship was found between the two variables indicating that students with lower affiliation needs tend to achieve more academically since they spend more time studying rather than focusing on their interpersonal relations (Harrell & Stahl, 1983). Thus it would follow that those students with higher affiliation needs would have low academic achievement. This view is in line with an earlier work that

states that people with a high need for affiliation compared to people with a high need for achievement neglect or de-emphasize achievement-related activities in favor of affiliative endeavors (Schneider & Green, 1977). However, Klein and Schnackenberg (2000) emphasized caution in examining the nature of the relationship of the two variables because environmental factors must be taken into account because individuals with high need for affiliation may also have increased performance if the work or task that they are supposed to do involves interaction with people. Related literature supports this since studies show that students whose need for warm and supportive relations tend to achieve better grades or perform better (Chan, 1980; Shu-Ping & Huang, 2016).

What is noticeably absent in the literature is the link between academic delay of gratification and need for affiliation. The researchers see this as a gap in the literature that needs to be clarified, hence this study.

Research Objectives

In line with the discussion above, the researchers believe that it is important to take a closer look at the variables and their relationships. The researchers put forth the argument that since literature has already established the relationship between the need for affiliation and academic achievement and academic delay of gratification and academic achievement, it is possible that the academic delay of gratification and need for affiliation are related. Similarly, the researchers put forth the concept that students who delay gratification in favor or academic tasks would have high academic achievement but low need for affiliation since the said students would value their academic goals over their need to connect with other people. In relation to this, the following research questions were formulated and answered.

1. Is academic delay of gratification related and a predictor of academic achievement?
2. Is need for affiliation related and a predictor of academic achievement?

3. Is academic delay of gratification related to need for affiliation?

Method

Participants and Design

A total of 1,021 students participated in this cross-sectional predictive study, ranging from 14 to 20 years of age. The participants were selected via convenience sampling, as the researchers were dependent only on the schedule and students assigned by their respective schools. They are fourth year high school students coming from different private and public high schools of selected cities in Metro Manila. All of the participants were Filipino, 602 of which were female (58.96%) and 419 were male (41.04%).

Measures

Academic Delay of Gratification Scale. This is a self-report measure constructed by Bembenutty & Karabenick (1998). The scale is a 4-point Likert scale instrument with ten items on measuring a student's academic delay of gratification. The reliability was established by Herndon (2008) with a Cronbach's alpha of .77.

Liking People Scale. This is a 5-point Likert scale instrument developed by Filsinger (1981). This 15-item instrument was developed for the purpose of measuring interpersonal orientation, the general liking of other people (Filsinger, 1981). From two samples of college students, it had an internal consistency of α=.85, and α=.75, respectively (Filsinger, 1981). In addition to the internal consistency, a coefficient alpha of .78 from the random sample of adults was established.

Procedures

The study was conducted in six different cities in the Metro Manila area, which includes the cities of San Juan, Manila, Mandaluyong, Marikina, Makati, and Quezon. The

research instruments were administered to the respondents in their respective classes based on the schedule given by the school principals. Following the general guidelines of ethical research, the researchers oriented and debriefed the participants before and after each administration. Average time for data gathering per class was somewhere between 20 to 30 minutes. All data were then scored and interpreted, sorted and readied for data analysis. During the data analysis phase, there was no attempt to categorize the data into public and private schools, as this is not within the scope of the present investigation. All data gathered were analyzed using descriptive statistics, zero-order correlation and regression analysis. All hypotheses were tested with .05 Alpha.

Results

Descriptive statistics were computed to determine the level of academic delay of gratification, need for affiliation and academic achievement of the respondents. Results showed that the high school students fall within the average level for all the research variables: academic delay of gratification (M= 31.82; SD = 4.97); need for affiliation (M = 50.01; SD = 9.38); and academic achievement (M = 85.14; SD = 4.61). SD values are also low indicating the homogeneity of the participants.

Zero-order correlation of the research variables

Data analysis showed that the academic delay of gratification is significantly and positively correlated with need for affiliation (r = 0.06, p< .05). Likewise, the relationship between academic delay of gratification and academic achievement is also significant (r = 0.26, p < .01) whereas the need for affiliation and academic achievement are negatively correlated with each other (r = -0.16, p< .01) confirming all the hypotheses put forth by the researchers.

Regression analysis

Multiple regression was performed to examine the predictive relations among the research variables with academic delay of gratification and need for affiliation serving as independent variables and academic achievement as the outcome variable. Results suggest both academic delay of gratification and need for affiliation significantly predicted the academic achievement of the respondents *(F*(2, 1018) = 55.804, $p < .01$, $R^2 = .099$, $R^2_{Adjusted} = 0.97$). Based on the analysis performed, the two independent variables can account for .97% of the variance observed in the academic achievement of the respondents. Looking at their respective beta weights (Table 1), the unique contributions of the predictors (.25 for academic delay of gratification and -.09 for need for affiliation) are all significant.

Table 1
Regression for Academic Delay of Gratification and Need for Affiliation on Academic Achievement

Predictors	**Academic Achievement**			
	B	SEB	*β*	*t*
Academic Delay of Gratification	.25	.03	.27	9.14*
Need for Affiliation	-.09	.02	-.18	-5.86*

*$p < .001$

Discussion

The present study looked into how the practice of academic delay of gratification and the one's need for affiliation can influence one's performance in school. Results showed that both variables are indeed a factor of academic performance wherein the more one practice academic delay of gratification the higher his or her chances of getting good grades. This is in line with the framework of Bembenutty and Karabenick (1998) that discussed the positive effects of

academic delay of gratification on grades. The positive effect of academic delay of gratification can be accounted for by its very nature as a dimension of self-regulation (Zimmerman, 1994;Eggen&Kauchak, 2012). When a student self-regulate as a function of his or her academics, all his or her cognitive behavior and affective resources are redirected from non-academic goals to the academic goals. Thus learners who forego immediate satisfying activities for the completion of academic tasks have a greater chance of getting a high grade. For instance, students who complete their homework and review first before watching television will most likely gain a higher score than those who engage first in the immediate gratifying acts.

In addition, since delaying gratification is associated with an orientation towards the future and thorough planning for temporally distant goals, students with a higher academic delay of gratification are more likely to pursue long-term goals, which can help in their academic success (Mischel, 1981). As such, it may then be viewed as a strategic asset by students for the attainment of more valuable rewards (Bembenutty & Karabenick, 1998).

Also consistent with previous literature (Harrell & Stahl, 1983), is the finding that the need for affiliation is negatively related to academic achievement. Similar to the academic delay of gratification, the discussion on the relationship and predictability of the need for affiliation is rooted in the nature of the variable. According to Harell and Stahl (1983), need for affiliation pertains to the need of people to prioritize and sustain interpersonal relationships resulting to the sacrificing of those things and opportunities that are irrelevant to this need. Relating the above framework to the current results, students acting on their need for affiliation could result to lower academic performance, as they tend to pick socializing over completing their academic tasks. In the age of social media, students may likely contact and chat with their friends rather than complete their homework or review their lessons or students may opt to attend a party or watch a movie over reviewing for an exam schedule the following day or go on vacation rather than prepare for an academic exercise.

The study also looked into the nature of the relationship of academic delay of gratification and the need for affiliation. Initially, the researchers hypothesized a negative relationship between the two variables but the data gathered shows otherwise. An analysis on the positive relationship of the variables would focus on the environmental misfits on a person's affiliation orientation and nature of academic task that was given to him or her (Klein &Schnackenberg, 2000) and this environmental misfit would dictate the motivation to complete a task. An example of an environmental misfit is when students with a high need for affiliation may have high academic delay of gratification due to the better fit that they have in the academic setting. Since the high school setting involves a lot of pair work or group work, students are motivated to perform academic tasks, which positively affect their ability to delay gratification. On the other hand, those with low need for affiliation may have low academic delay of gratification since they will be working in a scenario which they dislike; thus, leading to a lower motivation in fulfilling academic tasks and delaying gratification.

Conclusion

In summary, the present study puts forth the notion that the more a person delays gratification and complete academic tasks, the higher his or her academic score becomes while the more a person needs to affiliate with people, the lower his or her academic score becomes. Finally, the relationship between need for affiliation and academic delay of gratification may be affected by the fit of the person's need for affiliation and the nature of the task given; whether it requires a great amount of socialization or not. The present study is beneficial to the field of educational psychology because of the implications that the research variables have and how it affects students. Finally, the field of educational measurement may benefit based on the concept of environment fit where new measures based on that theoretical perspective maybe created and developed.

References

Arabzadeh, M., Kadivar, P., & Dlavar, A. (2012).The effects of teaching self-regulated learning strategy on students; academic delay of gratification. *Interpersonal Journal of Contemporary Research in Business, 4*(2), 580-587.

Bandura, A. (1986). From thought to action: Mechanisms of personal agency. *New Zealand Journal of Psychology*, *15*, 1-17.

Bembenutty, H. (1999). Sustaining motivation and academic goals: The role of academic delay of gratification. *Learning and Individual Differences, 11*(3), 233-257.

Bembenutty, H. (2007). Self-regulation of learning and academic delay of gratification: Gender and ethnic differences among college students. *Journal of Advanced Academics,18*(4), 586-616.

Bembenutty, H. (2011). Academic delay of gratification and academic achievement.*New Directions for Teaching and Learning, 126*, 1-124. DOI: 10.1002/tl.444

Bembenutty, H., & Karabenick, S. A. (1998). Academic delay of gratification. *Learning and Individual Differences, 10*(4), 329-346.

Bembenutty, H., & Karabenick, S. A. (2004).Inherent association between academic delay of gratification, future time perspective, and self-regulated learning. *Educational Psychology Review, 16*(1), 35-57.

Chan, R. M. (1980). The effect of student need for affiliation on performance and satisfaction in group learning. *Interchange on Educational Policy, 11*(1), 39-46.

Eggen, P., &Kauchak, D. (2012). *Educational psychology: Windows on classrooms* (9th ed.). USA: Peachpit Press.

Filsinger, E. (1981). A measure of interpersonal orientation: The Liking People Scale. *Journal of Personality Assessment, 45,* 295-300.

Harrell, A. M., & Stahl, M. J. (1983). Need for achievement, need for affiliation and the academic performance and career intentions of accounting students. *Journal of Accounting Education, 1*(2), 149-153.

Herndon, J. S. (2008). *The effects of delay of gratification and impulsivity on the academic achievement, substance abuse, &*

violent behavior of Florida middle-school and high school students in alternative learning settings (Doctoral dissertation). Retrieved from http:// http://accountability.leeschools.net/research_projects /pdf/StephanHerndon.pdf

Herndon, J. S., Bembenutty, H., & Gill, M. G. (2015). The role of delay of gratification, substance abuse, and violent behavior on academic achievement of disciplinary alternative middle school students. *Personality and Individual Difference, 86,* 44-49.

Hofer, J., & Busch, H. (2011). When the needs for affiliation and intimacy are frustrated: Envy and indirect aggression among German and Cameroonian adults. *Journal of Research in Personality, 45,* 219-228.

Klein, J. D., & Schnackenberg, H. L. (2000). Effects of informal cooperative learning and the affiliation motive on achievement, attitude, and student interactions. *Contemporary Educational Psychology, 25,* 332-341.

Mathieu, J. E. (1990). A test of subordinates' achievement and affiliation needs as moderators of leader path–goal relationships. *Basic and Applied Social Psychology, 11*(2), 179-189.

Mischel, W. (1981). *Introduction to personality* (3rd ed.). New York, NY: Holt, Rinehart and Winston, Inc.

Mischel, W. (1996). From good intentions to will power. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognitions and motivation to behavior* (pp. 99-129). New York: Guillford Press.

Pintrich, P. R., & De Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-50.

Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In D. H. Schunk & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 149-183). Hillsdale, NJ: Lawrence Erlbaum Associates.

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications*(7th ed.). New York, NY: The McGraw-Hill Companies, Inc.

Schneider F. W.,& Green, J. E. (1977). Need for affiliation and sex as moderators of the relationship between need for achievement and academic performance. *Journal of School Psychology, 15*(3), 269-277.

Schüler, J., Job, V., Fröhlich, S. M., & Brandstätter, V. (2008). A high implicit affiliation motive does not always make you happy: A corresponding explicit motive and corresponding behavior are further needed. *Motivation and Emotions, 32,* 231-242.

Shu-Ping, T., & Huang, C. Y. (2016). Effects of need for affiliation on performance and motivation in cooperative table tennis instruction. *International Journal of Humanities, Social Sciences and Education, 3*(1), 61-66.

Slabbinck, H., De Houwer, J., & Kenhove, P. V. (2012). The Pictorial Attitude Implicit Association Test for need for affiliation. *Personality and Individual Differences, 53,* 838-842.

Wiesenfeld, B. M., Raghuram, S., & Garud, R. (2001). Organizational identification among virtual workers: The role of need for affiliation and perceived work-based social support. *Journal of Management, 27,* 213-229.

Zhang, L., Karabenick, S. Maruno, S., & Lauermann, F. (2011). Academic delay of gratification and children's study time allocation as a function of proximity to consequential academic goals. *Learning and Instruction, 21*, 77-94.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81,* 329-339.

Zimmerman, B. J. (1994). Dimensions of academic self-regulation: A conceptual framework for education. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 3-21). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego, CA: Academic Press.

Educational Measurement and Evaluation Review (2016), Vol. 7 Issue 2

Test Development Using Differential Item Functioning

Arlene N. Mendoza
Pangasinan State University, Binmaley

Elsie M. Pacho
Don Mariano Marcos Memorial State University, Bacnotan

Abstract
Differential item functioning (DIF) analysis is an essential element in the evaluation of the fairness and validity of educational tests. This study developed a researcher-made test utilizing four DIF detection models: Mantel-Haenszel Chi-Square Statistic, Logistic Regression, Transformed Item Difficulty, and Rasch Model. Descriptive-comparative research design was employed in the DIF analysis based on students' differences on age, sex, language ability, socio-economic status, and school type. The study made use of the test scores of 188 BSE students major in Mathematics in the validated Achievement Test in Calculus I which was used as research instrument. Results of the study revealed that the revision and elimination of the potentially biased items in the test resulted to a valid, reliable, and fair test.Further, Mantel-Haenszel was the least sensitive in detecting DIF items among the models utilized. Moreover, the IRT Models, particularly the Rasch Model, revealed the highest number of detected DIF items, hence, has the highest statistical power of detection in the test constructed.

Keywords: Item bias, Development of an Achievement, Test, Rasch Model, Transformed Item Difficulty,Logistic Regression, Mantel-Haenszel Chi-Square Statistic

Introduction

In any assessment situations, one of the major goals of test developers is to ensure that the test instrument is free from bias against any identifiable groups. Bias is a major factor for tests considered unfair, inconstant, and contaminated by extraneous factors. A test is biased against or for a particular group if it under-predicts or over-predicts, respectively, their performance on the criterion of interest relative to some other groups (Pedrajita & Talisayon, 2009). Educational or psychological tests may include items that operate differently for certain groups. It is important to identify these items because they may lead to unfair results for groups being compared. The reason for such items to operate differently may be gender, age, culture, school type, teaching practices, classroom size, socio-economic status, or language differences between groups.

There are several methods of evaluating item bias, including the use of sensitivity reviews, differential validity studies, and Differential Item Functioning (DIF) detection methods (Wood, 2011). This study focused on item bias detection in an Achievement Test using differential item functioning (DIF) detection methodsfor test improvement. Differential item functioning (DIF) analysis is typically used to identify test items that are differentially difficult for respondents who have the same level of knowledge, skill, or ability but differ in ways that should be irrelevant to their performance on the test. The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with and items without DIF (Karami, 2012). Various differential item functioning (DIF) procedures have been proposed to assess potential bias. Despite the widespread application of DIF analysis in psychometric circles,it seems that the inherent complexity of the concepts in DIF analysis has hampered its wider application among less mathematically oriented researchers and only a limited number of them appearto be in current use. Thus, this study attempted to utilize these methods in assessing a

dichotomously scored test to detect bias test items and consequently construct a reliable, valid, and fair test.

This studyaimed to develop an Achievement Test by detecting biased test items particularly in Calculus usingItem Response Theory-based (IRT) model via Rasch Model and Transformed Item Difficulty Approach. In addition, two types of Classical Test Theory Models or the Contingency Table Approach via Logistic Regression and Mantel-Haenszel Chi-Square Statistics were also employed. According to Bradley (2009, p.5), "IRT techniques are the 'gold standard' of DIF detection." However, in the study of Salubayba (2013), she found out that Mantel-Haenszel and IRT-1PL were found both effective and sensitive in detecting DIF in the items. She showed that grouping variables like gender and school type were deemed to influence the performance of the pupils in reading comprehension and math application.However, in the study conducted by Madu (2012) to assess gender-related DIF using Transformed Item Difficulty, results show an incorrect picture of the quality of education for different groups and this may likely lead to the resources for education being distributed in an unfair manner. On the other hand, Pedrajita and Talisayon (2009) found out that there was a high degree of correspondence between the Logistic Regression and the Mantel-Haenszel Statistic in identifying biased test items. These findings gave the researchers an idea on applying an IRT-based models and Contingency Table Approach in detecting potentially bias item.

Further, comparativeanalyses among these DIF methods were done based on their sensitivity of detecting biased items. Moreover, the effect of biased items' elimination on the construct, content, and concurrent validity, and internal consistency reliability of the achievement test were determined.This study was delimited to some contextual variables such as age, sex, language ability, socio-economic status, and school type. In most situations, these factors were observed to affect examinees' chance to succeed in each test item.

The Achievement Test constructed focused on the topics about Calculus in order to construct set of valid,

reliable, and unbiased test items that wouldalso help dealing with problems monitoring the dynamical changes of biological samples, all kind of optimization problems or economic problems.Besides the significant aspect that this part of mathematics helps in development of an analytical mathematical thinking, calculus proves its effectiveness by solving real, practical problems. Calculus is used to find the rate of change; hence, it is very important because our society relies on it.

This study can significantly contribute to educational research especially in test development. Test experts, developers, and educators may: (1) gain insights on the applicability of DIF detection methods; (2) realize the validity of DIF methods in detecting biased test items based on students' differences on their age, gender, language ability, socio-economic status and school type; (3) use DIF methods in developing valid and equitable tests; and (4) employ DIF methods in purifying their assessment instruments.

Method

Research Design

This study employed the descriptive-comparative research design utilizing a researcher-made Achievement Test in Calculus. Development of the test was done by detecting item bias using the four methods of differential item functioning (DIF) models: Mantel-Haenszel Chi-Square Statistic, Logistic Regression, Transformed Item Difficulty, and Rasch Model. The DIF analysis of the test items were based on the students' differences on age, sex, language ability, socio-economic status, and school type. The detected biased items in the test using the four methods were revised and some were eliminated as based on the criteria set by the model. The validity and reliability of the test were then computed afterwards. The statistical power of detection of the DIF models was also determined through their sensitivity in detecting DIF items based on thesegroup

differences. The more DIF items detected, the higher the statistical power of detection of the DIF Models.

Participants

The test was administered to 188 college students taking up Bachelor of Secondary Education major in Mathematics from different Higher Education Institutions (HEIs) in Region I, private and public, who have already taken up their Calculus course, and enrolled during the first semester of SY 2014-2015.

Materials/Instrument

A questionnaire was formulated which solicited information regarding the students' age, sex, grade point average in Calculus I and in English I, socio-economic status, and school type. This information served as basis in detecting biased items. In addition, a researcher-made achievement test in Calculus I was constructed which consisted of 100 items. This is a multiple-choice test which covered concepts on Functions (8 items), Limits and Continuity (27 items), Derivatives (31 items), and Analysis of Functions and their Graphs (34 items).

Procedure

The researcher constructed an achievement test in Calculus Iand was evaluated by a panel of experts in the field of Mathematics who are at least Master's degree holder in Mathematics and have been teaching Calculus for at least five years. After validation, the test was administered to a group of BSE students major in Mathematics for field testing. The pilot testing has been conducted to a group of Bachelor of Science in Education majoring in Mathematics in Higher Education Institutions (HEI's) which were not included in the study.

When the test was tested for validity and reliability, it was administered again to 188 college students taking up

Bachelor of Secondary Education majorin Mathematics from different HEI's in Region I.The students were randomly assigned as the focal group and the reference group. The matched groups were based on the type of school they came from (public or private), their sex (male or female), their age (17 and below or 18 and above), grade point average in English I (above or below average of the group performance) and their socio-economic status in terms of gross monthly income (Php 8,000.00 and below or above Php 8,000.00). These groups were used as the bases in detecting DIF items through the DIF methods. The detected DIF items were then revised and improved if not eliminated. The revised version of the test was again subjected to test validity and reliability.The comparisons among the DIF methods were also done afterwards.

Data Analysis

This study has employed two Item Response Theory (IRT) DIF detection methods, the Transformed Item Difficulty approach, and the Rasch Model. Likewise, two Classical Test Theory (CTT) approaches were also considered, Logistic Regression, and Mantel-Haenszel Chi-Square Statistics. The efficacy of the methods was compared based on their sensitivity on detecting DIF items.

In calculating the MH statistics, the first step is to compute the probabilities of correct and incorrect responses for both groups. The second step is to find out how much more likely are the members of either group to answer correctly rather than incorrectly to the item. The overall DIF is calculated by summing the odds ratios at all ability levels and dividing them by the number of ability levels. The resulting index is the Mantel-Haenszel odds ratio denoted by α_{MH}. This index is usually transformed by the following: $\beta_{MH} = \ln\alpha_{MH}$ (Karami, 2012). A negative β_{MH} indicates DIF in favor of the focal group whereas a positive MH Δ shows DIF favoring the reference group (Wiberg 2007). Sometimes, β_{MH} is further rescaled into:

$$MHD = -2.35 \ln\alpha_{MH}.$$

A positive MHD indicates that the item was more difficult for the reference groups and a negative value shows that the focal group faces more difficulty with the item (Karami, 2012).

In Logistic Regression, an item is classified as displaying DIF if the two-degree-of-freedom Chi-squared test is beyond 5.9915 tested at 0.05 alpha significance and has a p-value less than or equal to 0.01 (set at this level because of the multiple hypotheses tested). Moreover, the Zumbo-Thomas (ZT) effect size measure had to be at least an R-squared of 0.130 (Zumbo, 1999). For ZT effect size measure, items were categorized as "A" if the value of their R-squared is significantly different from 0 and less than 0.13. Also, items were categorized as "B" if R-squared differ from 0.13 and less than 0.26. And it is considered under category "C" if R-squared differ from 0.26 and less than 1.

In Transformed Item Difficulty Approach, items with a perpendicular distance $(|D_i|)$values in excess of 1.5 reveal DIF. The larger (D_i) is, the more biased the item. A signed transformed difficulty measure of DIF, which preserved both the direction and magnitude of DIF was obtained by attaching a positive sign to (D_i) if the item reveals DIF in favor of the focal group, and a negative sign if the item reveals DIF in favor of reference group. For this study, a value of (D_i)greater than 1.5 indicates DIF,favoringthe focal group, whereas a value (D_i) less than -1.5 indicates DIF favoring the reference group.

The detection of differential item functioning through Rasch Model was performed using the Lord's chi-square method with one parameter logistic model. In this study, only one parameter was used ; hence, the Lord's chi-square with one parameter logistic model permits us to get item parameter estimates from the Rasch or one-parameter logistic (1PL) model. The calculated chi-square statistic was compared to a critical value (3.8415) based on an *a priori* specified level of significance (0.05), with degrees of freedom (df) corresponding to the number of parameters examined

for each item. If the observed chi-square exceeds the critical value, then the null hypothesis of no DIF is rejected.

Table 1 summarizes the detection threshold and the effect size of each DIF models in detecting DIF items.

Table 1
Detection Threshold and Effect Size of the DIF Detection Methods

DIF Detection Methods	Detection Threshold	Effect Size	Code	Scale Used
Mantel-Haenszel Chi-Square Statistics	3.8415	0.0 – 1.0	A	Delta Scale
		1.0 – 1.5	B	
		> 1.5	C	
Logistic Regression	5.9915	0.0 – 0.13	A	Zumbo and Thomas (ZT)
		0.13 – 0.26	B	
		0.26 – 1.0	C	
		0.0 – 0.035	A	Jodoign and Gierl (JG)
		0.035 – 0.07	B	
		0.07 – 1.0	C	
Transformed Item Difficulty	>1.5 and < -1.5	MHD value	N/A	N/A
Rasch Model	3.8415	0.0 – 1.0	A	Delta Scale
		1.0 – 1.5	B	
		> 1.5	C	

On the other hand, the validity and reliability of the test were determined using the following method:

a. The construct validity of the test was determined by showing that it is unidimensional. To evaluate unidimensionality, factor analysis was applied.

b. The concurrent validity evidence was secured by examining the relationship between predictors, which is the examinees' test score in the achievement test in Calculus I, and the criterion, which is the grade point average they obtained in their Calculus I course. Pearson Product Moment correlation coefficient is used to examine the relationship between the predictor and the criterion, and in this context the correlation coefficient is referred to as a validity coefficient (Reynolds et al., as cited in Pedrajita, 2009).

c. The content validity of the test was determined by computing a content validity index (CVI), using ratings of scale relevance by content experts. In this study, a 5-point rating agreement scale was used.

d. The internal consistency reliability of the original and the revised test versions was compared using the formula developed by Kuder Richardson, most commonly known as the KR-20. The KR-20 is sensitive to measurement error due to content sampling and is also a measure of item heterogeneity. It is applicable when test items are scored dichotomously, that is, simply right or wrong, as 0 or 1 (Reynolds et al., as cited in Pedrajita, 2009).

Results

I.Detection of Bias Items Using Differential Item Functioning Methods

A. Item Response Theory Models. The results disclosed that the DIF analysis through Transformed Item Difficulty approach detected more DIF items based on the examinees' differences on age. However, the level of sensitivity of the method on the examinees' differences on sex and socio-economic status was very low.The DIF items detected using Transformed Item Difficulty (TID) approach are consolidated in Table 2.

Table 2
Biased Items with Significant DIF across Matched Groups Using TID

Group Comparisons	DIF Items	Total
Age	1,2,3,4,6,7,8,9,11,13,14,16,20,25,27,28,30, 32,33,51,54,56,59,62,65,67,70,72,75,80,8 2,87,91,93,95,96,98	37
Sex	72	1
GPA in English I	6,8,16,27,55	5
Socio-economic Status	None	0
School Type	2,5,12,19,20,23,26,27,31,33,37,38,43,45,5 2,54,55,63,69,70,71,75,76,78,81,86,88,89, 92,95,100	31

Likewise, the Rasch Model DIF analysis also revealed the highest number of DIF items across students' age differences and small number of DIF items across sex and socio-economic status.The results of DIF detection analysis applying Rasch Model (RM) are presented in Table 3.

Table 3
Biased Items with Significant DIF across Matched Groups Using RM

Group Comparisons	DIF Items	Total
Age	1,3,4,7,9,10,11,13,14,15,17,18,19,20,22,23,24, 25,28,29,30,31,32,33,34,35,38,39,40,41,42,43, 44,47,48,49,50,51,52,53,54,55,56,57,58,59,60, 61,62,63,64,65,66,67,68,69,70,71,72,74,75,77, 78,79,80,81,82,85,86,87,88,90,91,92,93,94,95, 96,97,98,99,100	82
Sex	9,24,72,88	4
GPA in English I	7,9,10,11,15,19,22,23,26,33,40,52,53,55,56,61 ,62,67,71,72,74,76,77,78,86,88,90,91,93,97	30
Socio-Economic Status	74,86	2

School Type	1,3,6,7,15,21,23,33,34,35,36,37,41,43,46,47,52,54,58,59,61,62,64,67,69,70,71,72,74,75,88,92,93,95,96,97,99	37

B. Classical Test Theory Models. The findings on the DIF detection analysis using Mantel-Haenszel(MH) Chi-Square Statistic showed that the matched groups based on age differences flaggedmore DIF items. In comparison, results showed that MH Chi-square Statistic was also not very sensitive on detecting biased items in terms of examinees' differences on sex and socio-economic status as compared in the results obtained from using IRT Models. The findings are summarized in Table 4.

Table 4
Biased Items with Significant DIF across the Matched Groups Using MH

Group Comparisons	Identified DIF Items	Total
Age	8,10,11,13,15,16,19,22,34,37,44,52,53,57, 58,61,66,69, 72,74,77,81,96	23
Sex	37,72	2
GPA in English I	6,8,32,55,85,	5
Socio-economic Status	47,86	2
School Type	33,43,44,55,59	5

On the other hand, the Logistic Regression (LR) DIF analysis displayed highest sensitivity in terms of school type differences.In addition, the results showed that matching the examinees across sex and socio-economic status resulted to least number of detected DIF items. Table 5summarizes the DIF items detected using Logistic Regression DIF method.

Table 5
Biased Items with Significant DIF across Matched Groups Using LR

Group Comparisons	DIF Items	Total
Age	1,3,4,5,7,8,12,16,18,21,23,26,29,30,31,32,35,36, 37,38,40,41,45,46,47,48,50,51,60,63,67, 68,73,76,77,78,80,81,82,83,84,89,94,95,96,100	46
Sex	5,27,39,57,59,72,74,88	8
GPA in English I	6,8,16,17,22,26,53,55,61,66,71,73,74,76,79,85,9 0,93,95	19
Socio-economic Status	56,74,86,93,97,99	6
School Type	1,5,9,10,13,17,18,19,20,22,24,26,29,30,32,36,37 ,39,40,43,44,45,49,50,52,53,54,55,60,65,68,69,7 0,71,73,75,76,78,79,81,83,86,87,89,90,92,94,95, 100	49

II. Comparative Analysis on the DIF Detection Models

The comparative analysis of the four DIF methods applied to the 100-item dichotomously scored achievement test in Calculus Ifocused on the number of detected DIF items. The results of the detected DIF items by the four DIF methods are summarized in Table 6. The overall detection was based on the union of the detected items across all the matching variables.

Table 6
Detected DIF Items by the Four DIF Methods across Matched Groups

Matching Variables	Detected DIF Items (%)			
	MH	LR	TID	RM
Age	23	46	37	82
Sex	2	8	1	4
GPA in English I	5	19	5	30
Socio-Economic Status	2	6	0	2
School Type	5	49	31	37
Overall	32	82	60	89

Note. TID – Transformed Item Difficulty; MH – Mantel-Haenszel;LR – Logistic Regression; RM – Rasch Model

III. Validity and Reliability of the Revised Achievement Test

Based on the findings in the DIF analysis using the four DIF methods as well as in the validity and reliability analyses, the achievement test was revised. The revised test was composed of 50 items covering the four subtopics in Calculus I included in the test. It covered concepts on Functions (4 items), Limits and Continuity (13 items), Derivatives (16 items), and Analysis of Functions and their Graphs (17 items). Table 7 presents the final set of items included in the revised test.

Table 7
Items Included in the Revised Achievement Test

Topics	Items	Total
Functions	2,16,18,21	4
Limits and Continuity	26,27,31,33,34,35,36,38,39,41,42, 46,47	13
Derivatives	51,53,54,57,58,60,62,63,66,67,68,70,71, 73,74,75	16
Behaviors of Functions and their Graphs	22,23,24,5,8,10,15,44, 48,49, 77, 80,87, 94, 97, 98, 100	17

Further, the reliability and validity indices of the revised test are presented in Table 8. The table signifies that the revised version of the achievement test in Calculus I is valid, reliable, and a fair test. Thus, the test could be used in evaluating students' performance in Calculus I.

Table 8
Validity and Reliability Test of the Revised Achievement Test

Measures	Coefficient	Description
Construct Validity	0.667	Good
Concurrent Validity	0.159	Significant
Content Validity	0.9793 and 0.8965	Acceptable
Internal Consistency Reliability	0.822	Good

Discussion

I. Detection of Bias Items using Differential Item Functioning Methods

A. Item Response Theory Models. Table 2 shows that the highest number of detected DIF items in the Transformed Item Difficulty Analysis was observed across differences on age. This only indicates that this set of DIF items was not suited to the age level of one group. Hence, these items must be revised or replaced for further improvement of the test.

As gleaned further from the table, matching students in terms of socio-economic status does not detect any potentially biased items. This only show that the performance of the two different groups as based on their socio-economic status does not significantly varies in all the itemsincluded in the test. This finding only indicates that the test items were not bias against these groups. Hence, regardless of their status, the students could have the chance to succeed in all the items. Likewise, differences across gender detected only one DIF item. This also indicates that the students' chance on answering each item correctly in the test were not much affected by their gender differences except on item 72.

On the other hand, Rasch Model DIF analysis detected large number of potentially biased items when the examinees were grouped according to their age. This result states that the students' difference on age wasa great factor that could influence their probability of getting the correct answer to these 82 items. This finding further indicates that the set of items must be revised or replaced in order to suit to the level of ability of the disadvantaged group. Likewise, the finding connotes that the subject must be included in the curriculum of higher year level who are already prepared to take up this course.

Table 3 further indicates that the sensitivity of the Rasch Model in terms of gender and socio-economic status differences was very low. This only shows that thelevel of difficulty of the majority of test items was suited to the level

of ability of the students regardless of their sex and status in life.

B. Classical Test Theory Models. It is visible in Table 4 that the comparisons between groups of students of different age incurred the highest number of potentially biased items in the Mantel-Haenszel Chi-Square Statistic DIF analysis. On the contrary, the analysis detected few DIF items in terms of students' sex and socio-economic status differences. These results coincide with the findings obtained from the IRT Models. Also, this expresses that the sensitivity of the CTT and IRT in terms of age, sex, and socio-economic status differences weresomewhat comparable as based on the result of the test.

On the other hand, Logistic Regression DIF analysis shows that matching the examinees across school type, that is, private versus public HEI's, reveals the highest number of detected DIF items. This finding indicates that this factoralso affects the students' probability of succeeding on the 49 test items flagged with DIF.Hence, majority of the items are bias against school type.Thus, these items must be revised or replaced in order to suit to the capability of the students belonging to the disadvantaged group. Further, this result suggests necessary improvements in the educational system of the affected HEIs.

Moreover, Logistic Regression was also found to be less sensitive in DIF detection in terms of examinees' differences on gender and socio-economic status. The level of sensitivity of this method based on these two matching ability of the students is comparable to the previous three approaches. These results also connote that gender and socio-economic status differences are not contributors to the students' differing performances in the test and did not affect the students' chance of getting the given test items correct. In other words, the test items are not biased against these factors.

II. Comparative Analysis on the DIF Detection Models

The sensitivity of the four DIF methods in detecting DIF items were almost comparable as based on the result of DIF analysis across each matching variables. However, their sensitivity across all the matched groups differed as revealed by Table 6.

Table 6 discloses, that among the four DIF methods, Mantel-Haenszel Chi-Square Statistics detected the least number of DIF items. This implies that this method has the lowest statistical power of detection compared to the three methods. Hence, it is the least sensitive. This result confirms the study of Lopez (2012) which summarizes that the Mantel-Haenszel procedure is a straightforward and adaptable method for detecting DIF but this method has strong limitations which led to the development of other procedures.

On the contrary, Rasch Model appeared to be the most sensitive in the four DIF detection methods for having detected the highest number of items with DIF. This connotes that Rasch Model possesses the highest statistical power of detecting DIF items. Wiberg (2007) states that no matter which method is chosen, it is desirable that the method has high statistical power to detect DIF, that is, having high probability of identifying DIF in an item, while controlling for Type I error, which is the probability of identifying an item as DIF when the item has no DIF.

Moreover, between the two CTT-based methods, Logistic Regression was more sensitive compared to Mantel-Haenszel in detecting potentially biased items. However, it can be observed that the detection power of the Item Response Theory Model, particularly the Rasch Model, is higher than the Classical Test Theory Models. This only strengthens the findings that the latent score is a more precise measure of the ability of the test takers (Wiberg, 2007).

III. Validity and Reliability of the Revised Achievement Test

Construct Validity. The construct validity coefficients revealed that the revised version of the Achievement Test is a good test. Moreover, the results showed that the sampled test items in the revised test represent one dimension.

Concurrent Validity. The concurrent validity coefficient revealed that the revised version of the test obtained a positive relationship between the test score and the grade point average in Calculus I. Moreover, there exists a significant relationship between the two variables. This means that the revised version of the test is valid. However, this difference does not show any significance. This finding supports the results of Roznowski and Reith (1999) and Zumbo (2007) who have reported that DIF has little, if any, impact. Pae and Park (2006) however, reported that DIF may affect the performance on the test.

Content Validity. The content validity indices of the revised version of the test are within the acceptable level. Further, the results indicated that the test was judged valid by the evaluators.

Internal Consistency Reliability. The data in Table 8 revealed that the revised version of the test indexed a reliability coefficient greater than 0.8 which means that the set of test items are good and possess a reliable scale. However, the revised version has lesser reliability coefficient as compared to the original test. It can be observed that the test reliability coefficient obtained decreases when the number of items decreases. This result coincides with the results obtained in the study of Pedrajita (2007) which states that, as more responses on biased items were eliminated, the lower was the internal consistency reliability of the test version. Generally, the two tests were levelled as having a good internal consistency reliability; hence, they are comparable in terms of internal consistency reliability.

Overall, the results of the DIF item elimination on test validity show that the test is valid, reliable, and almost equitable for different types of examinees.

References

Bradley, K., et al. (2009). Constructing and evaluating measures: applications of the Rasch measurement model. *Application of Rasch Measurement,* University of Kentucky, Department of Educational Policy and Evaluation Studies. 131 Taylor Education Building, Lexington, KY 40506-0001.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59-76.

Lopez, G. E. (2012).Detection and classification of dif types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-sibtest, and logistic regression procedures. *Graduate School Theses and Dissertations*. Retrieved from http://scholarcommons.usf.edu/etd/4131.

Madu, B., (2012). Using transformed item difficulty procedure to assess gender-related differential item functioning of mulitple-choice mathematics items administered in nigeria. *Research on Humanities and Social Sciences, 2*(6), 41-56.

Pae T., & Park G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496.

Pedrajita, J. Q. (2007). Item bias elimination models for test reliability and validity. *Graduate School Theses and Dissertations*. UP Diliman College of Education: Philippines.

Pedrajita, J. Q., & Talisayon, V. M. (2009). Identifying biased test items by differential item functioning analysis using contingency table approaches: a comparative study. *Education Quarterly*, *67*(1), 21-43.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59,* 248–269.

Salubayba, T. (2013). Differential item functioning detection in reading comprehension test using mantel-haenszel, item response theory, and logical data analysis. *International Journal of Social Sciences, 14*(1), 76-82.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test. *Educational Measurement, 60,* 1-33.

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

Zumbo, B. D. (2007). Three generations of dif analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(*2*), 223-233. Lawrence Erlbaum Associates, Inc.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (dif): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human ResourcesResearch and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1997).A measure of effect size for a model-based approach for studying dif. *Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science*, University of Northern British Columbia: PrinceGeorge, B.C.

Educational Measurement and Evaluation Review (2016), Vol. 7 Issue 2

School Testing in the Philippines and the Need for Testing Standards and Guidelines

Violeta Valladolid
De La Salle University, Manila

Abstract

Tests and assessments are widely used for admission, placement, scholarship awards, psychological and educational screening of children with special needs, and career and vocational placement. Given their diverse and important use, measurement professionals have become increasingly concerned with questions of their validity, fairness, intended uses, and consequences. These issues have led to a dramatic increase in professional and technical standards on testing and assessment. The objectives of the study are two-fold: (1) to determine the current testing and assessment practices and procedures employed by the different educational institutions in the country; and (2) to come up with proposed guidelines or standards for school testing that are applicable in the Philippine setting. The study included two groups of respondents: (1) 50 personnel from 33 educational institutions who participated in the survey on the current testing and assessment practices and procedures, and (2) 15 who participated in a workshop on the development of school testing standards. The results of the study indicate that most schools have testing and assessment program that cater to the testing and assessment needs of their students. Tests are used for various reasons, particular for admission and for measuring students' ability, aptitude and attitude purposes. It is also good to note that the counselors and psychometricians handle the testing and assessment activities. Most of them also possess different kinds of tests that assess students' intelligence/IQ, aptitude, achievement, and personality. Seven school

testing standards and guidelines were proposed, which cover the following: (1) Test User Competence and Training, (2) Ethical and Professional Conduct of Test Users, (3) Test Selection, (4) Test Administration and Scoring, (5) Interpretation and Reporting of Test Results, (6) Rights of Test Takers, and (7) Use of Foreign Made Tests.

Keywords: testing and assessment, school testing standards, Philippine school testing

Introduction

Individuals are first exposed to tests and assessments very early in their school years. Tests and assessments are widely used in schools for admission, placement, scholarship awards, psychological and educational screening of children with special needs, and career and vocational assessment. In other countries, a number of mandated tests are administered to students, which compelled schools, teachers and students to put much time preparing for them. For example, a study by the Council of Great City Schools (2015) on the tests administered in 66 urban districts in the United States during SY 2014-2015 found that a typical student took 112.3 mandated standardized tests between pre-kindergarten classes and 12th grade. Students across grade levels spent on the average 4.8 hours (pre-K) to 25.3 hours (8th graders) during the school year taking the mandated assessments. This did not account for the quizzes or tests created by classroom teachers and the amount of time schools devoted to test preparation.

In Philippine schools, the Department of Education (DepEd) Order No. 55 s2016 (Department of Education, 2016) also provides that students take mandatory national tests to provide feedback on the current state of Philippine education and to assess the effectiveness and efficiency of the delivery of education services. The Early Language, Literacy and Numeracy Assessment is administered to Grade 3 students towards the end of the year while Exit Assessments are administered to the Grades 6, 10, and 12

students to determine if learners are meeting the learning standards of the elementary, junior high school, and senior high school curricula. The Career Assessment, on the other hand, seeks to determine the Grade 9 students' aptitude and occupational interests to guide them in their career choices.

The practice of testing and assessment in schools is not only limited to educational assessment. Educational assessment seeks to "determine how well students are learning … (and) provides feedback to students, educators, parents, policy makers, and the public about the effectiveness of educational services" (National Research Council, 2001, p. 1) and takes into account students' achievement, abilities, and learning outcomes. Psychological assessment is also widely used in schools to identify students who may have psychological, emotional, or behavioral difficulties. A psychological assessment is "an objective measure of samples of behavior including its causes, significance, and consequences. It may include the evaluation of social adjustment, emotional status, personality, cognitive/developmental functioning, language and information processing, visual-motor development, executive functioning, aptitude, academic achievement, and motivation" (Canadian Psychological Association, 2007, p.8). Results of educational and psychological assessments help the school, teachers, psychologists, and other concerned personnel to come up with academic and co-academic programs that will meet the student needs.

Given the diverse and important uses of tests and assessments, professional organizations come up with their own code of ethics that address the issues on competent assessment practice, test construction, and test use. According to Palladino Schultheiss and Stead (2008), ethical codes "describe a common set of principles and standards upon which practitioners can build their professional and scientific work" and "inform professional communities and societies about responsible assessment practices" (p. 604).

Existing Standards and Guidelines on Testing and Assessment

Ethical codes for psychologists have been developed by at least 71 national psychological associations across the globe (Leach & Harbin, in Palladino, Schultheiss, & Stead, 2008). The American Psychological Association (APA) was the first to adopt a formal code of ethics for any profession that uses assessments in 1952. Eighteen principles of this code addressed the issues on the use of psychological tests and diagnostic aids particularly on: (1) qualifications of test users (3 principles); (2) responsibilities of the psychologist sponsoring test use (4 principles); (3) responsibilities and qualifications of test publishers' representatives (3 principles); (4) readiness of a test for release (1 principle); (5) description of tests in manuals and publications (5 principles); and (6) security of testing materials (2 principles). Other organizations followed suit, which led to the increase awareness of the public on the appropriate use of tests and assessments (Camara, 1997).

At present, a number of testing standards and guidelines are in place to guide testing practitioners in the conduct of educational and psychological testing and assessment. The International Test Commission (ITC) has developed guidelines on adapting tests (2005), test use (2001), computer-based and internet-delivered testing (2005), quality control in scoring, test analysis and reporting of test scores (2012), security of tests, examinations, and other assessments (2014), and practitioner use of test revisions, obsolete tests, and test disposal (2015). The Joint Committee on Testing Practices has developed the *Code of Fair Testing Practices in Education* (2004), which is a "guide for professionals in fulfilling their obligation to provide and use tests that are fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics" (p.1). The American Federation of Teachers, the National Council on Measurement in Education, and the National Education

Association jointly developed the *Standards for Teacher Competence in Educational Assessment of Students* (1990). The Joint Committee on Testing Practices (JCTP), which was established in 1985 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) has developed The *Standards for Educational and Psychological Testing (2014)*, *Code of Fair Testing Practices in Education (2004), Responsible Test Use: Case Studies for Assessing Human Behavior (2010),* and *Assessing Individuals with Disabilities in Educational, Employment & Counseling Settings (2002),* among others.

These above standards are universal and internationally-accepted and in fact, are encouraged to be made as benchmarks or the basis from which to develop locally applicable documents as these will promote a high level of consistency across national boundaries (International Test Commission, 2001).

Testing in Philippine Schools

A number of issues and concerns face the Philippine schools with regard to testing and assessment. First, the twin problem of the inapplicability of foreign-made tests and the dearth of locally-made tests has been recognized as early as the 1970's. Guanzon (1985) also noted the tendency of test users to use foreign-made tests *in toto,* without attempting to adapt these tests, through test or item modification, test translation, or development of local norms.

Most Philippine schools also do not have comprehensive testing and assessment program. One of the primary reasons for the inability to provide such program is the unavailability of standardized psycho-educational tests as well as of experts or personnel to conduct the assessment in the schools. Referral to outside agencies or psychologists or experts also poses a problem since psycho-educational testing is very expensive, and thus, not affordable especially for parents and children from poor families. Schools, outside agencies, psychologists, and experts/practitioners also rely heavily

on foreign-made standardized tests. This again would pose some problems since, aside from too costly, these tests are being questioned with regard to their validity and applicability for use in other cultures (Valladolid, 2014).

The DepEd mandated-tests also did not skip the ire of some organizations or associations related to education. For example, the Federation of Association of Private Schools and Administrators (FAPSA) has called on the Department of Education (DepEd) to abolish the National Achievement Test (NAT), saying, "students need to think, not memorize." FAPSA President Eleazardo Kasilag said that schools teach only what students are most likely to encounter during exams, such as in Science, Math, English, Filipino and Sibika, making students abandon assignments that require critical thinking in favor of drill, memorization, and repetitive practice (Flores, 2014).

To regulate the practice of psychology and psychometrics in the Philippines, the Republic Act No. 10029, known as the "Philippine Psychology Act of 2009", was signed into law in March 2010. The law aims to protect the public from inexperienced or untrained individuals offering psychological services. Article III, Section 3b provides that the practice of assessment by licensed psychologists covers diverse types of clients consists of delivery of psychological services that involve among other things, psychological assessment, particularly in

> *gathering and integration of psychology-related data for the purpose of making a psychological evaluation, accomplished through a variety of tools, including individual tests, projective tests, clinical interview and other psychological assessment tools, for the purpose of assessing diverse psychological functions including cognitive abilities, aptitudes, personality characteristics, attitudes, values, interests, emotions and motivations, among others, in support of psychological counseling, psychotherapy and other psychological interventions* ("An Act to Regulate", 2010, p. 3).

This means that the scope of psychological assessment also covers that of assessment in the educational context especially in assessing cognitive abilities, aptitudes, attitudes, values, interests, emotions, and motivations. This defined role implies a collaborative work between assessment/testing practitioners and psychologists (Magno, 2010). The passing of the law needs trained and licensed practitioners, particularly psychologists and psychometricians, to handle testing and assessment activities in school and industrial settings.

The professionalization of the testing profession demands the need for standards and guidelines that will guide the schools, test users and test takers in the Philippines. This is to encourage best practice in the field of testing and assessment and to prevent their negative consequences and misuse.

Objectives of the Study

The objectives of the study are two-fold: (1) to determine the current testing and assessment practices and procedures employed by the different educational institutions in the country; and (2) to come up with proposed guidelines or standards for school testing that are applicable in the Philippine setting.

Method

The study included two groups of respondents: (1) 50 personnel from 33 educational institutions who participated in the survey on the current testing and assessment practices and procedures, and (2) 15 who participated in a workshop on the development of school testing standards that will guide schools, test users, and test takers in the Philippines. The data from the two surveys were analyzed using frequencies and percentages. Responses to open-ended questions were content analyzed. Table 1 presents the profile of the participants.

Scope and Limitation of the Study

This study is a preliminary study to get an overview of the testing and assessment practices in selected Philippine schools. It involved a purposively selected participants of the International Conference on Educational Measurement and Evaluation (EME) held in September 2016 chosen as they are practitioners of EME in their schools. Data were drawn from a survey form analyzed using descriptive statistics, which the researcher acknowledge to be limited, hence recommends a more comprehensive study in the future.

Table 1
Profile of the Respondents

Profile	f	%
A. Survey on Testing and Assessment Practices (N=50)		
Gender		
Male	13	26.00
Female	37	74.00
Total	50	100
Educational Background		
Bachelors	23	46.00
Masters	22	44.00
Doctoral	5	10.00
Total	50	100
Place of School		
Metro Manila	32	64.00
Rizal	3	6.00
Luzon	15	30.00
Total	50	100
Job Position/Title		
Faculty/Teacher	22	44.00
Dean/VP/Executive Head	5	10.00
Psychometrician	8	16.00
Office Head/Officer (Testing, Assessment, etc.)	4	8.00
Evaluation Asst.	3	6.00
Admissions Coordinator	1	2.00

Math Coordinator	1	2.00
Staff Assistant	1	2.00
NR	5	10.00
Total	50	100

B. Survey on School Testing Standards (N=15)

Gender		
Male	4	26.67
Female	11	73.33
Total	15	100

Results

The results of the study are presented following the sequence of the objectives of the study indicated in the earlier part of the study.

Testing and Assessment Practices and Procedures

Availability of Testing/Assessment Program. It is good to note that of the 33 educational institutions, majority (27 or 81.80%) of the respondents indicated that their schools have their own testing and assessment program. Only five (15.20%) indicated having no testing/assessment program (Table 2).

Table 2
Distribution of Schools with Testing Assessment Program

Availability	**f**	**% (N=33)**
With Testing/Assessment Program	27	81.80
Without Testing/Assessment Program	5	15.20
NR	1	3.00
Total	33	100

Inability of the school to develop a testing/assessment program (80%) and unavailability or lack of experts or personnel to handle the testing program

(60%) are the top-most reasons given by the respondents for the lack or absence of an assessment system in their schools. Two (40%) respondents cited the high cost of test materials as the other reason. (Table 3)

Table 3
Reasons for Not Having Testing/Assessment Program/Center

Reasons	f	% (N=5)
School was not able to develop a testing program.	4	80.00
There is no personnel/expert to handle the program.	3	60.00
Tests are too expensive.	2	40.00
There is no space or room to house testing center.	0	0
The school does not see the need for it.	0	0

The most popular means by which these schools get information about their students' ability, aptitude, personality, and attitude is through the use of other forms or methods of assessment and by engaging the services of outside agencies or psychologists (Table 4).

Table 4
Other Ways to Get Information on Students' Ability, Aptitude, Personality, or Attitude

Ways	f	% (N=5)
Engages the services of outside agencies/psychologists	3	60.00
Uses other forms/methods of assessment	3	60.00

Nature of Schools' Testing and Assessment Program

Purpose of Testing and Assessment Activities. All (100%) respondents whose schools have testing

program indicated that they conduct testing and assessment for student admission purposes. Others use tests to determine their students' ability, aptitude, and attitude (22 or 81.48%), as supplement or support to their counseling services (18 or 66.67%), and for career and placement services (18 or 66.67%). A little half of them indicated using tests to identify students with learning or special needs (15 or 55.56%) and to screen personnel applicants (14 or 51.85%).

Table 5
Purposes of Testing/Assessment Activities

Purposes	**f**	**% (N=27)**
For admission purposes	27	100
To determine students' ability, aptitude, attitude, etc.	22	81.48
To supplement or support for counseling	18	66.67
For career and placement services	18	66.67
To determine who has learning or special needs	15	55.56
For personnel applicant screening	14	51.85
For referral to outside testing	8	29.63

Types or Forms of Assessment Used. Majority of the respondents indicated that their schools rely on their counselors (81.48%) and psychometricians (70.37%) to handle the testing and assessment of their students. Eight (29.63%) schools each also seek the services of psychologists and classroom teachers. Only two respondents cited hiring consultants.

Table 6
Personnel Handling the Assessment of Students

Personnel	f	% (N=27)
School Counselors	22	81.48
Psychometricians	19	70.37
Psychologists	8	29.63
Classroom teachers	8	29.63
Consultants	2	0.07
Others		
-CEM outsource	2	0.07
- principal	1	0.04
- statistician	1	0.04

Both foreign-made and locally-made tests are employed by the schools, as indicated by 21 (77.78%) respondents each. Other schools, however, rely on school-developed tests (59.26%), teacher-made tests (55.56%), and DepEd tests (37.04%).

Table 7
Types or Forms of Assessment Employed by Schools

Types	f	% (N=27)
Foreign-made standardized tests	21	77.78
Locally made tests	21	77.78
School-developed tests	16	59.26
Teacher-made tests	15	55.56
DepEd tests	10	37.04

Various tests are being used by the schools, which include IQ or intelligence (74.07%), aptitude (81.48%), achievement (66.67%), and personality tests (66.67%). There are also some schools that make use of reading tests (48.15%) and performance-based tests (40.74%). Only a few of them use projective tests (7 or 25.93%).

Table 8
Types of Test Used by Schools and Outside Agencies/Psychologists

Response	f	% (N=27)
IQ/intelligence tests, such as	20	74.07
Aptitude tests	22	81.48
Achievement tests	18	66.67
Personality tests	18	66.67
Projective tests	7	25.93
Reading tests	13	48.15
Performance-based tests	11	40.74

Aside from using tests to assess students' cognitive abilities and personality, the respondents also indicated that their schools make use of other methods and approaches. The most-commonly used method or approach is reviewing of child's school records and past evaluation results (85.19%). Direct observations, projective tests, and interviews (77.78%), task analysis or assessment of the student's works (62.9%), and questionnaires/checklist for parents, teachers and students (59.26%) are also employed.

Table 9
Other Methods/Approaches Used to Assess Students

Types	F	% (N=27)
Review of child's school records and past evaluation results	23	85.19
Assessment of student works (task analysis)	17	62.96
Direct observations, projective tests, and interviewing of child	21	77.78
Use of parent, teacher, and student questionnaires/ checklist	16	59.26
Interview with parents	0	0

Development of Guidelines and Standards on School Testing

There are a number of guidelines and codes of ethics that were formulated by various organizations. While these are applicable to and can be used in the Philippine context, there is still a need to come up with guidelines relevant to the specific needs and conditions in the Philippine schools. As such, the proponent formulated some ethical and professional standards and guidelines for school testing that will guide schools, test users, and test takers in the country. This is also to encourage best practice in the field of testing and to prevent the negative consequences of its misuse.

Proposed Standards/Guidelines for School Testing

Seven school testing standards and guidelines were proposed, which cover the following: (1) Test User Competence and Training, (2) Ethical and Professional Conduct of Test Users, (3) Test Selection, (4) Test Administration and Scoring, (5) Interpretation and Reporting of Test Results, (6) Rights of Test Takers, and (7) Use of Foreign Made Tests.

These proposed standards are not an invention of new guidelines but they will represent the previous work of specialists and organizations on psychological and educational testing standards. The aim of this set of standards is to bring together the common concepts and principles that embody existing guidelines, standards, codes of ethics, and other related documents, and to come up with guidelines that are much needed by Philippine schools and industrial setting.

Proposed Guidelines on Test Users/ Practitioners Competence and Training. The lack of adequate training and experience of testing personnel in many Philippine schools is one concern that needs to be addressed. Thus, it is advisable that they update their skills and competencies by either enrolling in post-graduate courses or by attending relevant seminars and

conferences. In addition, it is necessary that guidelines be developed to help them in the execution of their assessment tasks and responsibilities. Some of these guidelines cover the need for them to have formal academic coursework in tests and measurement, to use tests that they are only competent to administer and interpret, to observe the classification system and requirements and qualifications specified by test publishers, and to participate in continuing education.

Proposed Guidelines on Ethical and Professional Conduct of Test Users. Test users are expected to maintain and adhere to the highest standards of ethical practice. They should act in professional and ethical manners and should treat all people involved in the testing process with respect. The proposed guidelines cover the need for test users to set and maintain high personal standards of competence, to only offer testing services and use tests for which they are qualified, to purchase only tests that they are qualified in accordance with the competency levels set by test producers, to ensure that test materials are kept confidential, to protect tests from unauthorized access, and to respect copyright law and agreements stipulated in the test manual.

Proposed Guidelines on Test Selection. Test users should be able to justify the selection of tests to be administered to the client. Tests selected should be appropriate for the client's needs and status, purpose of testing, and setting (i.e., school or industry). This can be done by reviewing the test manual and materials to be able to get complete information about the test, selecting tests based on the appropriateness, evaluating the evidence of technical quality (i.e., validity, reliability) of the test, ensuring the availability of evidence that the tests have validity to predict performance in another situation or setting, and evaluating samples of test questions to check the appropriateness of the test.

Proposed Guidelines on Test Administration and Scoring. Test users are expected to be able to use

the prescribed and standardized procedures to administer and score the test. The proposed guidelines cover the need for the test users to secure informed consent from the client or his/her parents before test administration, to follow the established or standardized procedures for administering in a standardized manner as stipulated in the Test Manual, to conduct the test administration in a structured and controlled environment, and to protect the security of test materials.

Proposed Guidelines on Interpretation and Reporting of Test Results. Test users should be able to ensure that test data are interpreted appropriately, also taking into consideration the limitations of test data. They should also ensure that information from testing is not misused. They should release the interpreted results only to those who have a legitimate right to receive that information. Such guidelines would include the need for test users to follow the procedures in scoring and interpreting the results as set in the Test Manual, to avoid making generalizations about the test takers based on a single test score, to gather other information about the test taker to support the interpretation of test results, and to communicate the test results in a timely fashion and in a manner that is understood by the test taker. These guidelines also require that test users to consider multiple factors that can compromise a test taker's performance on the test, to develop norms based on the intended population of test takers, to ensure that test results are presented in a form that is understandable to the recipient, and to ensure that only qualified personnel will receive the raw data on his/her behalf.

Proposed Guidelines on Rights of Test Takers. Test takers who have the highest stake in the whole testing process also have rights and responsibilities and they should be informed about them. Test takers have the right to know in advance about the purpose and use of testing, coverage of the test, and the types of question included. They should also be told how to get information to help them or their parents/guardians judge whether the test

should be taken, how long scores will be kept on file, to whom the results will be released, and in what manner test scores and related information will or will not be released. Test takers should also be treated with courtesy, respect, and impartiality, regardless of their age, gender, religion, SES, and other personal characteristics. They should also be assured that interpreted results are released only to those who have a legitimate right to receive the information.

Proposed Guidelines on the Use of Foreign-Made Tests. Because of the lack of locally-made psychological and educational tests, there is a proliferation in the use of foreign-made standardized tests, not only in schools but also in industrial setting for employment testing. However, validity is one of the most important attributes of a good assessment. Validity is commonly referred to as the extent to which a test measures what it purports to measure. It also reflects the degree to which interpretations of the test scores are valid reflections of the skill or proficiency that an assessment is intended measure (Pitoniak, Young, Martiniello, King, Buteux, & Ginsburgh, 2009). Using foreign-made tests may bring about threats (i.e., construct-irrelevant variance) that are irrelevant to the skills or proficiencies being measured. One reason for having guidelines on using foreign-made tests is to minimize these threats to validity and to maximize the degree to which the test scores reflect the true ability level of the test taker in the content area being assessed and minimize the impact of other factors on test scores, such as level of English language proficiency, educational background, and cultural background and differences.

The International Test Commission (ITC) has formulated the "Guidelines for Translating and Adapting Tests" in 2005, covering four different categories: Context Guidelines, Test Development and Adaptation Guidelines, Administration Guidelines, and Documentation/Score Interpretation Guidelines. Most relevant of these and which can be adapted to the Philippine setting are the guidelines in test administration.

The proposed guidelines will cover the need for

test users to be sensitive to a number of factors related to the test materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores, and to ensure that tests to be used are unbiased and appropriate for the various groups that will be tested. They also require test users to review if the constructs being assessed by the test are meaningful to the test takers, to check if the test includes evidences and empirical studies on possible group differences in performance on the test, to ensure that there is validity evidence to support the intended use of the test for other groups not included in the normative data, to conduct validity study to ensure the applicability of the test to the intended test takers, and to come up with norms based on the intended population.

Discussion of the Results

The results of the study indicate that most schools have testing and assessment program that will cater to the testing and assessment needs of their students. Tests are used for various reasons, particular for admission and for measuring students' ability, aptitude and attitude purposes. It is also good to note that the counselors and psychometricians handle the testing and assessment activities, since they normally have academic background in test and measurement. Most of them also possess different kinds of tests that assess students' intelligence/IQ, aptitude, personality, achievement, and personality.

Schools also employ other ways to meet their assessment needs and requirements. These include review of the student's school records and past evaluation results, interview of parents, assessment of student works, and direct observation of students' behavior. This is true most especially for schools that do not have testing and assessment program.

However, while the result may mirror or depict, to some extent, the testing and assessment practices in big and private schools, we cannot ascertain if this is also the reality in other parts of the country and most especially in

small schools and public schools across the country. The cost of standardized tests, the lack of qualified personnel or expert to handle, the lack of funds/budget, and the failure of the schools to realize the importance of a testing program are some of the factors that may hinder schools from putting up a testing program. As such, a more in-depth and extensive study should be conducted involving more schools and of different types (public vs. private schools), grade or year levels offerings (basic vs. secondary vs. higher educational institutions), and setting (urban vs. rural areas). Other aspects of testing and assessment practices also need to be explored, such as, the types and specific tests used and for what purposes, qualifications of personnel handling the testing and assessment in terms of academic qualifications and years of experience, quality or effectiveness of test administration, scoring, interpretation and reporting of test results, and the practice of assessing students with special needs.

The importance of test and assessment necessitates the need for standards and guidelines that will guide the schools, test users and test takers in the Philippines. This is to encourage best practice in the field of testing an assessment and to prevent negative consequences and misuse of testing. First, having testing and assessment standards that are applicable to the Philippine milieu is needed to ensure that the test user has the necessary competencies to carry out the testing process, and the knowledge and understanding of tests and test use. This is also to protect the test takers, who have the greatest stake in the testing activities.

Second, the proposed guidelines will help schools to be aware of the consequences of using foreign-made tests that were developed in different settings using different samples as normative group. This will also encourage schools to develop tests that are applicable to their own clientele or students.

Third, since the role of testing and assessment specialists is increasing due to the demand for quality assurance in schools, especially in teaching and

implementation of programs and due to the shift from national testing to institutional testing to support instruction, research and organizational performance (Magno & Gonzales, 2011), guidelines such as these will provide the schools parameters on how school testing should be conducted in their schools.

The proposed standards/guidelines need to be further studied and validated by experts and practitioners in the field of psychological and educational testing and assessment. It is hoped that the final sets of testing standards will be developed and eventually adopted by schools and organizations that are engaged in school testing and assessment.

References

An Act to Regulate the Practice of Psychology Creating for this Purpose a Professional Regulatory Board of Psychology, Appropriating Funds Therefore and for Other Purposes. (2010). Retrieved from http://psych.upd.edu.ph/downloadables/ra10029.pdf

Camara, W. J. (1997). Use and consequences of assessments in the USA: Professional, ethical and legal issues. *European Journal of Psychological Assessment, 13*(2), 140–152.

Canadian Psychological Association. (2007). *Professional practice guidelines for school psychologists in Canada: The CPA section of psychologists in education.* Ontario, Canada: Canadian Psychological Association.

Carlota, A.J. (1980). Research trends in psychological testing. In A. Carlota & L. Lazo (Eds.), *Psychological measurement: A Book of readings (*pp. 31-47). Quezon City: UP Psychological Foundation.

Code of Fair Testing Practices in Education. (2004). Washington, DC: Joint Committee on Testing Practices.

Council of Great City Schools. (2015, October). *Student testing in America's great city schools: An inventory and preliminary analysis.* Retrieved from

http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

Department of Education. (2016, June 30). *Policy guidelines on the National Assessment of Student Learning for the K to 12 Basic Education Program.* Retrieved from http://www.deped.gov.ph/sites/default/files/DO_s2016_55.pdf

Flores, H. (2014, March 11). Private schools seek abolition of achievement test. *Philippine Star*. Retrieved from http://www.philstar.com/headlines/2014/03/11/1299486/private-schools-seek-abolition-achievement-test

Guanzon, M.A. (1985) "Paggamit ng panukat na sikolohikal sa Pilipinas: Kalagayan at mga isyu". In A.Aganon & M.A. David (Eds.), *New directions in indigenous psychology: Sikolohiyang Pilipino, isyu, pananaw at kaalaman* (pp. 341-362). Manila: National Bookstore.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing,* 1(2), 93-114.

Layton, L. (2015, October 24). Study says standardized testing is overwhelming nation's public schools. *Washington Post.* Retrieved from https://www.washingtonpost.com/local/education/study-says-standardized-testing-is-overwhelming-nations-public-schools/2015/10/24/8a22092c-79ae-11e5-a958-d889faf561dc_story.html

Lazo, L. S. (1977). Psychological testing in schools: An assessment. *Philippine Journal of Psychology, 11*(1), 23-27.

Lazo, L. S., de Jesus-Vasquez, M.L., & Tiglao, R.E. (1975). A survey of psychological measurement in the Philippines: Clinical, industrial and educational settings. In A. Carlota & L. Lazo (Eds), *Psychological measurement: A book of readings* (pp. 2-30). Quezon City: UP Psychology Foundation.

Magno, C. (2010, July). A brief history of educational assessment in the Philippines. *Educational Measurement and Evaluation Review,* 1, 140-149.

Magno, C., & Gonzales, R. DLC. Measurement and

evaluation in the Philippine higher education: Trends and development. In E. A. Valenzuela (Ed.), *UNESCO policy series: Trends and development in Philippine education* (pp. 47-58). Philippines: UNESCO National Commissions.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington DC: National Academy Press.

Palladino Schultheiss, D. E., & Stead, G.B. (2008). Ethical issues in testing and assessment. In J.A. Athanasou & R.V. Esbroeck (Eds.), *International handbook of career guidance* (pp. 603-62). Netherlands: Springer.

Pitoniak, M. J., Young, J. W., Martiniello, M. King, T.C., Buteux, A. & Ginsburgh, M. (2009). The *Guidelines for the assessment of English-language learners.* Retrieved from https://www.ets.org/s/about/pdf/ell_guidelines.pdf

Standards for Teacher Competence in Educational Assessment of Students (1990). Retrieved from http://buros.org/standards-teacher-competence-educational-assessment-students

The Standards for Educational and Psychological Testing (2014). Retrieved from http://www.apa.org/science/programs/testing/standards.aspx

Valladolid, V.C. (2014). *Evaluating the validity of the dual discrepancy model in identifying students at-risk of reading disability in Philippine public schools.* (Unpublished dissertation). De La Salle University, Manila.