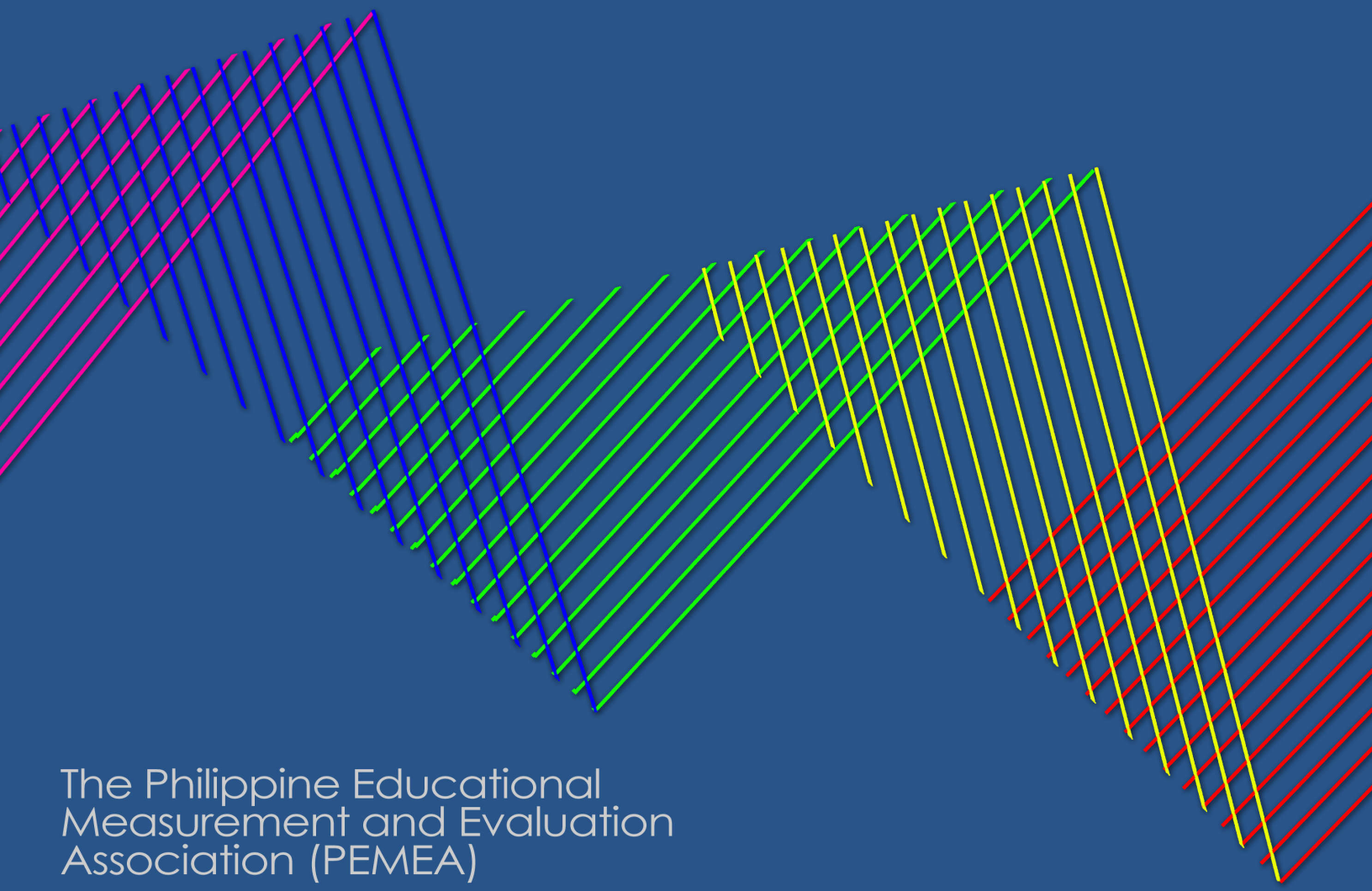
ISSN 2094-5876

# The Educational Measurement and Evaluation Review

VOLUME 2 JULY 2011



The Philippine Educational Measurement and Evaluation Association (PEMEA)

The Educational Measurement and Evaluation Review is one of the official publications of the Philippine Educational Measurement and Evaluation Association (PEMEA). The EMEReview publishes scholarly reports about contemporary theory and practice in the field of education and social science that highlights measurement, assessment, evaluation, psychometrics, psychological testing, and statistics. The journal is international, refereed, and abstracted. The journal is presently abstracted in the Asian Education Index, Social Science Research Network, Google Scholar, Open J-Gate, and NewJour.



The articles in the EMEReview are open access at http://pemea.club.officelive.com/EMEReview.aspx

Publication Division of PEMEA
Philippine Educational Measurement and Evaluation Association

**Educational Measurement and Evaluation Review**
ISSN 2094-5876 http://pemea.club.officelive.com/EMEReview.aspx ◈ July 2011, Vol. 2

A Perspective in Educational Measurement: An Editorial Note
*Carlo Magno* ..... *1*

## Empirical Reports

Investigating the Development of Analytical Skills in Teacher Education
*Anders Jonsson & Sven A. Lennung*..... *3*

Temperament Styles of Children from Samoa and the United States
*Carmelo M. Callueng, Desmond M. Lee Hang, Richard DLC. Gonzales*
*Ainslie Chu Ling-So'o, and Thomas D. Oakland* ..... *18*

Monitoring Teacher Trainees' Mathematical Competence in an Accelerated Teacher Education Program
*Karoline Afamasaga-Fuata'i* ..... *35*

Determining Experts and Novices in College Algebra: A Psychometric Test Development and Analysis Using the Rasch Model (1PL-IRT)
*Jonathan V. Macayan and Bernardino C. Ofalia* ..... *62*

Exploring the Factors of Perfectionism within the Big Five Personality Model among Filipino College Students
*Joel C. Navarez and Ryan Francis O. Cayubit*..... *77*

Using Logistic Regression and Mantel-Haenszel Statistic in Differential Item Functioning Analysis: A Comparative Study
*Jose Q. Pedrajita*..... *92*

Attributes of Distracters of Multiple Choice Items with Potentially Biased Options
*Janet Lynn S. Montemayor* ..... *111*

"We're classmates, can we be friends?": Translation and Validation of the Filipino Version of Classmates' Friendship Questionnaire (CFQ) in the Philippines
*Fraide A. Ganotice, Jr. and Jonalyn B. Villarosa* ..... *124*

Exploratory and Confirmatory Factor Analysis of Self-efficacy among Student-Athletes
*Maria Cristina M. Firmante* ..... *137*

Beyond assessment: Impact Evaluation of a Community-Based Education Development in Lao PDR
*Benjamina Gonzalez-Flor, Richard DLC Gonzales, and Alexander Gonzalez Flor* ..... *148*

## Short Report

A Book Review on "Designing Written Assessment for Student Learning"
*Karina M. Agustin* ..... *171*

Educational Measurement and Evaluation Review (2011), Vol. 2, 1-2


http://pemea.club.officelive.com/ EMEReview.aspx

# A Perspective in Educational Measurement: An Editorial Note

**Carlo Magno**
*PEMEA Board Member*
*EMEReview Editor*
*De La Salle University, Manila*

The second volume of the Educational Measurement and Evaluation Review (EMEReview) presents a perspective in the practice and theory of measurement and evaluation in the educational setting. I am pleased to inform the readership about some of the accomplishments of the journal. First, is the increased citation of the journals' first volume and the indexing in several databases that is searchable in the internet. The journal is now included in search databases across different universities in the world. Second, is the participation of international experts in the editorial board of the journal that includes John Hattie (The University of Auckland, New Zealand), Jack Holbrook (University of Tartu, Estonia), Anders Jonsson (Malmo University, Sweden), Timothy Teo (University of Auckland, New Zealand)), Tom Oakland (University of Florida, USA), Jimmy dela Torre (Rutgers University, USA), Jose Pedrajita (University of the Philippines, Diliman), Shu-ren Chang (Department of Testing Services, American Dental Association, USA), and Karma El Hassan (Office of Institutional Research and Testing, American University of Beirut, Lebanon). Having the wide participation of editors and reviewers around the world brings competitiveness of the journal with others in the same field. A wide array of perspective is provided for the papers given the participation of reviewers in the international scene.

The second volume brings 10 studies and one article. These articles selected for this volume provide empirical studies about the current advancement of educational assessment in the world. Authors from different parts of the world showed a reflection about the state of educational measurement in the field.

The first article by Anders Jonsson and Sven A. Lennung assessed the development of students' analytical skills that includes observation, analysis, and taking action. Their analysis provided a perspective on the need to improve teacher education that allows students to become analytical in different context and subject areas. The second article by Carmelo Callueng, Desmond M. Lee Hang, Richard Gonzales, Ainslie Chu Ling-So'o, and Thomas D. Oakland differentiated four temperament styles (using the Student Styles Questionnaire) by varying age, gender, race/ethnicity, geographic region, and school type in categories. They found that the differences in the data can be explained by drawing cultural and social characteristics of the samples studied. The third article by Karoline Afamasaga-Fuata'i investigated the effectiveness of the Accelerated Diploma in Education Program (ADEP) on teacher trainees' mathematical performance. The tests assessing mathematical performance were analyzed using the Rasch model. The author through the generated item maps was able to compare item and person distributions within tests. The fourth article by Jonathan

Macayan and Bernardino Ofalia also applied the one-parameter Rasch model for a College Algebra Diagnostic Test. The Rasch model was used to determine person and item reliability, standard errors, item difficulty, and dimensionality of the test. The fifth article by Joel Navarez and Ryan Francis Cayubit further validated the construct perfectionism through the Big Five personality traits. The correlation through a structural model provides a perspective that personality traits can be a good criterion when perfectionism is used as a predictor. The sixth article by Jose Pedrajita used both logistic regression and Mantel-Haenszel statistics to detect Differential Item Functioning (DIF) of a Chemistry Achievement Test for junior high school students. Certain recommendations were given for the potentially biased items that were determined. The seventh article by Janet Lynn Montemayor analyzed the distracters of Research Competency Test for Graduate Students. She also differentiated the distracters across private and state universities, and students living in highlands and lowlands through DIF analysis. The eighth article by Fraide Ganotice, Jr. and Jonalyn Villarosa developed a Filipino version of the Classroom Friendship Questionnaire (Miscenko & Rascevska, 2008). The factor structure of the adapted scale was proved to have good validity. The ninth article by Maria Cristina Firmante developed a measure of self-efficacy in sports for student athletes. After conducting a principal components analysis to uncover the factor structure of the scale, another sample was used to confirm the measurement model. The tenth article by Benjamina Gonzalez-Flor, Richard Gonzales, and Alexander Gonzalez Flor conducted an impact of a community-based intervention on educational outcomes. The indictors the educational outcome includes enrolment, promotion rates, repetition rates, gender parity, and completion rates. The last article by Karina Agustin is a book review by Magno and Ouano's "Designing Written Assessment for Student Assessment." The book was assessed in the areas of structure, inclusion of a software in the package, approach and presentation of the book, clarity and organization, and content.

The articles in this present volume cover practices on testing, measurement, assessment, and evaluation. The evaluation studies and assessment reports in this issue serve as models that institutions need to undertake to determine the effectiveness and quality of delivery in their programs. This calls for a greater use of assessment and evaluation that informs decision.

Educational Measurement and Evaluation Review (2011), Vol. 2, 3-17


http://pemea.club.officelive.com/ EMEReview.aspx

# Investigating the Development of Analytical Skills in Teacher Education

**Anders Jonsson & Sven A. Lennung**
*Malmö University*

**Abstract** An important aspect of teacher competence is analyzing complex classroom situations and suggesting appropriate actions that follow from the analysis. Novice teachers' analyses are, however, typically simpler than analyses done by experienced teachers. The aim of this study was to investigate whether the analytical skills of pre-service teachers had developed throughout teacher education, and whether the pattern of strengths and weaknesses in students' performances during their first semester had changed at the time of graduation. The results show that the skills did not improve during teacher education, since the students performed at the same level during the first and the last semester. Only two changes where identified: (1) The students had *increased* their awareness about the need to have more information before being able to make well-grounded decisions; (2) The students considered different motives for acting in particular ways *to a lesser extent* during the last semester.

**Keywords:** assessment, competence, evaluation, teacher education

## Introduction

An important aspect of teacher competence is the ability to analyze complex classroom situations, such as being able to identify a problem, figure out what caused the problem, understand students' incentives for behaving in particular ways, and also to take appropriate actions following this analysis. There are a number of studies, however, which report that beginner and novice teachers' analyses of classroom situations are typically simpler and more descriptive than are analyses done by more experienced teachers (e. g. Berliner, 1986; Carter, Cushing, Sabers, Stein, & Berliner, 1988; Lin, 1999). This raises the question of how pre-service teachers can acquire this expertise, and a typical answer is that it requires extensive experience, while teacher education is usually seen to have little impact on such skills. On the other hand, arguments have been made about the possibility to feed forward this learning process, claiming that experiences need not necessarily be made by the students themselves, but can also be made vicariously, for instance through role plays or simulated situations (Elliott, 1991; Metcalf, Ronen Hammer, & Kahlich, 1996). Further, there is both theoretical and empirical support for the assumption that students' learning can be enhanced by making expectations explicit to them (Black & Wiliam, 1998; Frederiksen & Collins, 1989; Sadler, 1989).

Following this latter set of arguments, an assessment methodology called the "Interactive examination", where students analyze classroom situations simulated through digital video, was developed in order to assess, as well as to support, pre-service

teachers' learning of analytical skills. Previous research has shown that this "Interactive examination" can indeed be considered a valid instrument for assessing preservice teachers' analytical skills (Jonsson, Baartman, & Lennung, 2009; Jonsson, Mattheos, Svingby, & Attström, 2007), and that this methodology supports student learning and improves their performance (Jonsson, 2010; Jönsson, 2008). This article reports on a comparison between students' results on this examination during their first and last semester respectively, aiming to investigate whether the analytical skills of pre-service teachers developed throughout the teacher-education program, and whether the pattern of strengths and weaknesses noted in student performance during their first semester changed.

## Background

Competence, as the concept is used here, refers to the integration of knowledge, skills, and attitudes into situation-relevant actions, in order to master relevant tasks (Taconis, Van der Plas, & Van der Sanden, 2004). To be "competent" thus means to be able to *act knowledgeably* in relevant situations. This definition suggests that no matter how much you know, you cannot be considered competent unless you can actually use this knowledge to solve problems within a certain field of practice. Moreover, it means that competence is not something we are born with, but rather a quality which can be learned and improved.

This definition of competence, however, raises the question of how beginners turn into competent workers. There have been several answers to this question, with contributions from prominent authors such as Gilbert Ryle, Michael Polanyi, and Donald Schön. However, the novice-to-expert framework presented by the Dreyfus brothers (1986) and the theory of "legitimate peripheral participation" by Jean Lave and Etienne Wenger (1991) are perhaps the most widely cited in the literature on progression from novice to competent. In the context of teacher education, much work has been done by David Berliner and his associates, referring mainly to the Dreyfus framework. One of the things that Berliner turns our attention to is that competent teachers can identify, analyze, and act upon things that go more or less undetected by novices. For example, in a study by Sabers, Cushing, and Berliner (1991), teachers with varying experience and expertise in teaching viewed three different television monitors. Each monitor focused on a group of junior high school students, and the participants had to express their thoughts as they viewed the monitors. They also had to answer questions about classroom management and instruction. What could be seen in this investigation was that the teachers categorized as "experts" were able to monitor, understand, and interpret events in more detail, and also with more insight, than the participants categorized as either "novices" or "advanced beginners." Further, the teachers differed in how they attended to the "multidimensional nature of the classroom." As the authors express it: "Experts not only used all three monitors to view the classroom activities, but by all outward appearances they seemed more at ease in completing the task. They gave the overall impression of enjoying the experiment and participated enthusiastically" (Sabers et al., 1991, p. 76). Similar results are presented from other studies (e.g. Berliner, 1986; Carter et al., 1988; Lin, 1999). In addition, these results are not confined to the "holistic recognition of patterns" in experts (see e.g. Berliner, 2004, p. 207) or tacit knowledge; pre-service teachers also have difficulties in dealing with general and theoretically-grounded issues when reflecting on teaching events (van den Berg, 2001).

Given the conclusion that novices and experts differ in their ability to identify, analyze, and act upon multifaceted information in the classroom, how then can the students be guided towards this competence? Sabers et al. (1991) argue that there is probably a limit to what can be learned in teacher education, and that it takes extensive time to acquire competence in such a complex domain. This argument is supported by Björklund (2008) in a review of research on experience-based learning, where he suggests that the difference between problem solving by experts and novices can be explained by the fact that experts have acquired a larger knowledge base of "implicit memories" (i.e. memories that are not consciously attained) than the novices. In principle, experts have in some sense "seen it all before," and can therefore act in an intuitive and non-reflective manner. Unfortunately, for teacher educators this would mean that there is no way to fast forward this process, and that "teacher education can, at best, start people on the path toward expertise and provide them with the tools and dispositions to better learn from their experience" (Sabers et al., 1991, p. 85).

As Elliott (1991) notes, however, even though professional learning (just like any other learning) is situated and experiential, this does not mean that is has to involve direct participation. Practical situations can also be experienced vicariously, for example by reflecting on case studies and/or discussing different ways to act in relation to simulation exercises. That case studies can indeed be effective in this regard is shown in a study by Metcalf et al. (1996). Here a group of students were exposed to a series of campus-based activities, which included role play and giving short lessons which were videotaped. The students also watched simulated classroom situations on video. With these situations as starting points, they had to provide explanations, suggest possible solutions, and propose potential consequences of the situations. Every activity was analyzed and discussed by the students in groups. The "reflective ability" of these students was then compared to a group of students who had been exposed to regular field-based education. Results from this comparison showed that the group of students exposed to campus-based activities had significantly improved their skills in identifying critical events in complex, pedagogical situations. They could also give more advanced explanations to these situations and were more inclined to provide rationales for different actions taken. Even skills in carrying out meaningful lessons, as measured in this study, were improved by this group, whereas the skills of the control group had not changed. The results from Metcalf et al. thus suggest that, by working systematically with simulated situations (such as video sequences and role play), pre-service teachers can (1) improve their skills in identifying critical events in complex situations, (2) give more advanced explanations to these situations, (3) become more inclined to give rationales for actions taken, as well as (4) improve their performance in giving lessons.

## Assessing Competence to Support Student Learning

If we once again, in the light of the discussion above, ask the question: "Given the conclusion that novices and experts differ in their ability to identify, analyze, and act upon multifaceted information in the classroom, how can the students be guided towards this competence?," we might re-phrase the answer. If students can learn from others' experiences – through the use of role plays, case studies, or simulations – and if we (through the use of language) can communicate the standards of quality performance to the novices, then they would not have to work it out all by themselves in the course of their own experiences. Well-designed instruction in combination with explicit criteria and standards might in this way help the novices to focus on relevant details in the performance of others, and thus potentially make their own experiences

more effective, as well as making it possible for students to self-assess their own performance.

These assumptions underlie an assessment methodology called the "Interactive examination," originally developed by Mattheos, Nattestad, Falk Nilsson, and Attström (2004) for dental students, which in this study has been adapted to teacher education. The examination consists of several parts, but only those relevant to the current study will be described here. For a more comprehensive report, see Jonsson et al. (2007) or Jönsson (2008). The examination is performed over the Internet, and one of the main parts is a personal task in the form of an authentic, professional problem, presented to the students as a classroom case, simulated through digital video (cf. Metcalf et al., 1996). Each student watches three cases and for all movies the students must: (1) Describe the situation without prejudice (Observation), (2) State a problem and analyze the situation displayed (Analysis), and (3) Formulate what actions should be taken, considering the needs of all those involved (Taking action). Students then submit their answers as word-processed documents. Since teacher competence involves handling a wide variety of situations, ranging from providing appropriate conditions for student learning to attending to an individual student's social and psychological difficulties, from assessing student knowledge, to arranging meetings with students, parents, and/or colleagues, etc., the personal tasks in the "Interactive examination" have to reflect a similar complexity for the examination to provide valid data about student performance. Therefore the tasks do not focus on details or well-defined problems. Instead, they are (more or less) open for interpretation, so that the students themselves have to choose what is important, and identify one or more problems to be solved.

When submitted, students' answers are assessed with the help of a scoring rubric. To achieve as high reliability as possible, as well as to aid in giving detailed feedback, the rubric used is analytic (i.e. designed for assessing different aspects individually) rather than holistic (Jonsson & Svingby, 2007). For each of the "global questions" (Observation, Analysis, Taking action) there are four or five assessment criteria in the rubric, resulting in a rubric with a total of 13 criteria (the different aspects assessed by the rubric are listed in Table 1). This means that the rubric was designed to capture what the students saw in the movies, how they interpreted the situations, and what strategies they used to deal with the situations displayed. The rubric was distributed to the students approximately three weeks before the examination, so that they could read and discuss the criteria with peers and instructors. The purpose of sharing the criteria with the students is that, in order for the assessment to support student learning, it had to be clear to the students what was expected of them (see e.g. Black & Wiliam, 1998). This sharing of criteria is one of the important differences between this study and the studies by Metcalf et al. (1996) and Mattheos et al. (2004) discussed above.

The "Interactive examination" was developed with the dual purpose of assessing the acquisition of selected aspects of teacher competence, as well as supporting student learning of these same competences, following the call from Frederiksen and Collins (1989) that assessments should lead to students becoming more skilled at whatever the assessments are set out to measure. The validation for both summative and formative purposes has been investigated and reported elsewhere (Jonsson et al. 2009; Jonsson, 2010; Jönsson 2008), showing for instance that the inter-rater reliability is reasonably high (Spearman's rho .795; $p<.01$), although the tasks are open ended, and that students perceive the examination to be very authentic and relevant to their future profession.

To summarize: Competent teachers seem to be able to identify, analyze, and act upon things that are not always noticed by novices and also provide analyses that are more insightful, for instance going beyond simple questions of classroom management. In order to develop these skills it has been suggested that novice teachers need to gain extensive experience as teachers and that teacher education therefore might have a very limited influence on students' professional development. As outlined above, however, well-designed instruction in combination with explicit criteria and standards might help novices to focus on relevant details, making it possible for students to improve their analyzing performance without extensive personal teaching experience. This has been shown to be the case in a previous study (Jonsson, 2010), where pre-service teachers' analyses were greatly improved by providing the students with scoring criteria and exemplars. These findings suggest that at least certain analytic and reflective skills might be developed during teacher education, with the aid of vicarious experiences and clear expectations.

In the current study, the performance of students who carried out the "Interactive examination" during their first semester of the teacher-education program is compared to the performance of the same students during their last semester. The aim is to investigate (1) if the students' analytical skills have developed throughout the teacher-education program, and (2) whether the pattern of strengths and weaknesses noted in student performance has changed.

## Method

### Context

The teacher-education program investigated in this study consists of three different components: a Course in the beginning and at the end of the program, covering general areas in the teaching profession that are common to all teachers regardless of subject major; A major subject, including both content knowledge and pedagogical content knowledge; and one or more minor subjects that do not need to include pedagogical content knowledge.

As part of the program, students are also assigned to "partner schools," where they are supposed to participate in the day-to-day activities during their school-based education and, of course, learn how to teach through real-life experiences. Notable is that there is no single course, or set of courses, in which the practicum periods are gathered. Instead, the practicum periods are always "integrated" with campus-based education into courses, encompassing learning of both practical and academic nature. The actual level of integration might vary, however, as the use of quotation marks above indicates. A specific problem for instance is that the teacher educators observing the students during the practicum are not necessarily experts in the same subject as the students are teaching, which might lead to a focus on general issues (such as classroom management), since the educator is not familiar with the particular pedagogical content knowledge of the subject taught. Another problem that might affect the level of actual integration is that the assessment of student performance is often clearly divided between campus-, and school-based education, where the former primarily focuses on subject-related knowledge and the latter on procedural skills (e.g. Hegender, 2010). In relation to the particular focus of this study, there is a risk that intellectual skills may "fall between two stools" in the current organization, since they are neither clearly subject-related nor clearly procedural. This means that, unless there has been any intervention specifically aimed at developing analytical skills (besides the "Interactive

examination"), the students are left to develop the ability to analyze complex classroom situations by themselves on an experiential and intuitive basis.

## Sample

The results of this study is based on a small sample of pre-service teachers (n = 19) specializing towards teaching in primary school. In 2004 a number of pre-service teachers ($N$ = 171) carried out the "Interactive examination" during their first semester of the teacher-education program, and in 2007 those students specializing in teaching in primary school were asked to take the examination once more during their last semester. Of the 61 students specializing in teaching in primary school, and who carried out the "interactive examination" in 2004, 19 students agreed to take the examination once more during their last semester (response rate 31 %). The reasons for not taking part in the study are not known, but the small sample does not differ significantly from the initial sample with respect to the scores in the 2004 examination. In Sweden, the teaching in secondary school program is one year longer than for teaching in primary school. This means that those students specializing in teaching in secondary school were scattered throughout the university at this moment, taking different subject courses depending on their different majors, and were not included in the study.

## Research Data and Analyses

Data for this study consists of two sets of answers from the personal task (i.e. classroom situations simulated through digital video) in the "Interactive examination", one from 2004 and one from 2007, for the same students. In each of these data sets, students responded to the "global questions" (Observation, Analysis, Taking action) for three different movie sequences. There is considerable variation in length for students' answers, but a typical answer may cover slightly less than a page, which means that the total material amounts to approximately 100 pages of student writings.

Student answers were assessed with a scoring rubric (the skills measured by the instrument are shown in Table 1) and the assessments were carried out by external assessors. Since the rubric has three levels (Fail, Acceptable, and Excellent), which were assigned 0, 1, and 2 points respectively, and each student did three movies, the range of scores for each criterion in the rubric was 0-6.

Since the (dependent) sample was not normally distributed, Wilcoxon's signed-rank test, which is a non-parametric analogue to the t-test, was used both for the sample as a whole, comparing students' examination scores from their first and last semester, as well as for individual students, comparing the change in personal scores. In the latter case, the sub-scores for each criterion in the rubric formed the basis for the analysis (i.e. the sum of the scores from the three movies for each of the 13 criteria). Further, frequency analyses were done, investigating how well the students performed in relation to each of the individual criteria in the rubric. From this information, the pattern of strengths and weaknesses in students' answers from the first and last semester were compared, in order to identify qualitative differences in their analyses.

## Different Conditions

Although most conditions were similar, there were some notable differences between the examinations in 2004 and 2007 respectively. First, in 2004 the students could choose three movies from a pool of nine movies. The movies in this pool

differed with respect to the age of the children, the situations displayed, and the subject context. In 2007 there were only three movies that all students had to analyze. Two of these movies were new to the students, but one had also been used in 2004 and had therefore been analyzed by the most of the students. Second, in 2004 the students had access to the scoring rubric during the examination. The rubric, by making explicit what was to be assessed, could be used as a "tool of thought," aiding the novice students when performing their analyses, in order to support student learning and improve their performance. However, at the end of their education, the students should hopefully manage without such scaffolding structures, and therefore the students in 2007 did not have access to the rubric. Third, the assessment was carried out by different assessors in 2004 and 2007. However, as reported in Jonsson et al. (2009), both internal consistency of the assessors (Cronbach's alpha .793 and .833 respectively) and the inter-rater reliability have been shown to be quite high (.902; $p<.001$, for the overall score and .795; $p<.01$, at the criterion level), suggesting that although the use of different assessors will undoubtedly contribute to unwanted variance, this variation is small as compared to variance due to students' performance.

## Results

The results show that, when analyzed as a group, there was no significant difference in students' scores on the examination when comparing scores from their first and last semester. The mean difference in total score was less than 1.5 points (maximum 78 points) between 2004 and 2007. When analyzing individual scores a somewhat different image emerges. Of the 19 students, nine had improved their scores, nine had lowered them, and one had exactly the same score. Most of these changes were small, however, and only two of the students significantly altered their scores ($p<.05$). On the other hand, these two students made very large improvements to their scores: 71 and 37 percent respectively on the total score.

When looking at changes in relation to individual criteria, there are some criteria where students showed greater difficulties in 2004 than in 2007, but also some where the reverse is true. Again, most of these differences are small. There are, however, some notable deviations from this general picture. In 2004, seven of the 19 students had difficulties (defined as failing to comply with the lowest standard in the rubric in at least one of the three movies) in discussing conceivable motives for the behaviors shown by students and teachers in the movie sequences. Three years later, in 2007, almost all of the students (16) had difficulties with this criterion.

A criterion where students are assessed as to whether they can specify what additional information is needed in order to make a well-grounded decision demonstrates the very opposite. Here all but one student had difficulties in 2004, while none had problems with this in 2007. There are also two criteria, where most students score very low in both 2004 and 2007. The first is when they are supposed to discuss possible consequences of the situations in the movies. The second weakness is their ability to use research-based arguments to justify their suggested actions. These changes in the pattern of strengths and weaknesses noted in student performance are summarized in Table 2.

Table 1
*The aspects of analytical skills assessed by the rubric in the "Interactive examination"*

| *Observation*<br>The student should be able to: |
|---|
| Describe the situation without prejudice |
| Focus on relevant details |
| Describe the perspectives of all those involved in the situation |
| Describe the situation so that other people can understand it |
| *Analysis*<br>The student should be able to: |
| Identify those involved in the situation |
| Identify a problem |
| Interpret the situation: Why did this situation occur? |
| Discuss conceivable motives for the behaviors shown |
| Discuss conceivable consequences of the situation |
| *Taking action*<br>The student should be able to: |
| Give suggestions as to what additional information is needed in order to make a decision |
| Suggest actions that take both students' and the teacher's perspectives into account |
| Suggest alternate actions to be taken |
| Discuss pros and cons with different actions, both in a short and a long term perspective |
| Suggest actions that are in line with the observation and the analysis made |
| Justify the actions suggested by making references to course literature or other relevant sources |
| Widen the discussion by not only focusing on the context shown in the particular situation, but also include societal aspects such as curriculum, culture, different social categories, etc. |

Table 2
*A summary of the changes in the pattern of strengths and weaknesses noted in student performance between 2004 and 2007*

| Criterion | Number of students displaying difficulties[1] in | |
|---|---|---|
| | 2004 | 2007 |
| The analysis discusses conceivable motives for the behaviors shown. | 7 | 16 |
| Gives, when relevant, suggestions as to what additional information is needed in order to make a decision. | 18 | 0 |
| The analysis discusses conceivable consequences of the situation. | 12 | 19 |
| The actions suggested are supported by references to course literature or other relevant sources. | 18 | 17 |

*Note.* 1. Defined as failing to comply with the lowest standard in the rubric for at least one of the three cases.

## Discussion

The aim of this study was to investigate whether the analytical skills of pre-service teachers specializing in teaching in primary school could be demonstrated to have developed throughout the teacher-education program, and whether the pattern of strengths and weaknesses noted in students' performances during their first semester had changed.

### Have the Students Developed their Analytical Skills?

The results show that, when analyzed as a group, there was no significant difference in students' scores on the examination when comparing scores from their first and last semester. Furthermore, of the 19 students investigated, only two significantly improved their scores. Before jumping to the conclusion that the teacher-education program has failed in helping the students to develop these skills, some methodological issues need to be considered. First, the group of students investigated in this study is but a small self-selected sample (31 %) of the students that took the examination in 2004. There was, however, no significant difference between this small sample and the larger 2004 sample with respect to the scores on the 2004 examination.

Second, while the students in 2004 had access to a scoring rubric, the students in 2007 did not have anything to guide them while analyzing the movies. The purpose of this arrangement was to see whether the students – after going through the entire teacher-education program – had developed the "intellectual tools" needed in order to analyze complex classroom situations. In this way, the knowledge gained through their teacher education was supposed to compensate for the more explicit guidance provided by the rubric. As can be seen in the results, this actually happens, since the last-semester students do perform at the same level as the first-semester students, but without the help of the rubric.

Third, the different assessors in 2004 and 2007 might have given rise to unwanted variability. However, this variation is thought to be small as compared to variance due to students' performance.

Another issue that may have affected the results is that one of the movies in 2007 was already used in the 2004 examination. Since most of the students analyzed this movie back then, they might have remembered how to analyze it, and in this way affected the results. However, no significant difference in students' scores could be found between this movie and the other (new) movies.

Last, but not least, as opposed to the students in 2004, the students in 2007 were volunteers, who did not have any stakes (in the form of grades) in the assessment. This means that they had no external incentives to do well on the examination, which might have affected their willingness to engage in the tasks. However, the fact that they did volunteer, and also that they completed all tasks, might reflect an intrinsic motivation and interest in the competences assessed.

Taken together, since the students performed at the same level during their last semester as during their first, but without the assistance of the rubric, a very optimistic conclusion is that the students have developed some "tools of thought" to apply when analyzing complex classroom situations. In this respect they can be said to have developed the analytical skills to some extent throughout the teacher-education program, even if their results on the "Interactive examination" have not changed.

A more realistic, although pessimistic, interpretation might focus on the fact that last-semester students performed at the same level as they did as first-semester students, implying that there has not been any substantial development during the teacher-education program. Even if the first-semester students were aided by a rubric, which the last-year students were not, the rubric was designed for assessing first-year students, and it might be argued that you could expect more from last-year students. If the students had developed their analytical skills further, ceiling effects could be expected when using a rubric designed for first-semester students. This, however, was not the case in the current study. A problem here is that (in Sweden) it is only mandatory to provide the goals (i. e. "expected learning outcomes") in the course syllabus, but not any standards. This means that while it could be stated for instance that the students are expected to be able to "discuss the influence of different social and cultural conditions" after a specific course in the teacher-education program, there is also not necessarily any information on *how well* they are supposed to "discuss the influence of different social and cultural conditions." Consequently, there are no commonly accepted standards to compare the last-year students' performances with, to see whether there has been sufficient development during the program. Still, it is obvious from this study that there has been very little progression from the first to the last semester and that these students are still very much novices when it comes to analyzing complex classroom situations. This finding would seem to lend support to the notion of Sabers et al. (1991), that teacher education has limited influence on these competencies. It should be acknowledged, however, that there have been no specific instructional interventions directed towards these competences, except for the "Interactive examination" during the first semester, included in the campus-based education for these students. Further, as was discussed previously, there is a risk of excluding these intellectual skills in the school-based education, since the assessment often seems to focus on procedural skills (Hegender, 2010).

## Has the Pattern in Students' Performances Changed?

Even if there has been no significant change in students' overall scores from the first to the last semester regarding their analytical skills, there can still be changes according to individual criteria. When comparing the pattern of strengths and

weaknesses in students' performances, however, the overall picture is also that of no change. Most students in 2007 managed to cope with the criteria that the students succeeded in complying with in 2004. Also, the weaknesses noted in many of the students' analyses in 2004 (i.e. not being able to discuss the consequences of the situations displayed and not being able to theoretically justify their suggested actions), were still present at the time of graduation.

There were some differences, however. For instance, the students have learned to judge whether there is enough information available in order to make a well-grounded decision, which was a weakness in many of the students' analyses during the first semester. Although it cannot be ascertained that this skill was acquired through teacher education, it is quite plausible considering the fact that none of the students mastered this skill in the beginning of the program, while most of them did in the end. This indicates that the students have become aware of the need to have more nuanced information about the (school) students and the context in order to plan for action to be taken, as opposed to suggesting ad hoc solutions based on the (possibly scarce) information available.

There were also differences in the opposite direction, where the students performed less well in 2007 than they did in 2004. This may seem a bit puzzling, since we are not dealing with factual knowledge that might well be lost in the haze of time, but with more consistent skills that are probably not as easily forgotten. What these results suggest is therefore not necessarily that the students have lost the ability to discuss conceivable motives for the behaviors that students or other teachers display (which is the criterion in question), but that this quality is not part of the "tools of thought" for analyzing complex classroom situations that the students have developed during their education. Either they have not encountered or thought of this particular aspect, which means that they do not know that it is missing in their analyses, or they have considered it unimportant and consciously left it out. Since being able to see the situations through the eyes of other persons (either students or teachers), and trying to understand the reasons for their behaviors, could be seen as a very important skill of teachers, it seems unlikely that the students would consciously leave this aspect out of their analyses. A more likely interpretation is that they do not automatically consider different reasons for other peoples' behaviors when analyzing situations like these, but make assumptions about the reasons for the behaviors displayed, possibly based on own experiences. Needless to say, this limits the number of potential solutions seen and suggested by the students and may also make these solutions less appropriate for individuals not sharing the same experiences as the pre-service teachers.

A similar argument can be made for not being able to discuss conceivable consequences of a situation, as it seems unlikely that the students would consciously leave this aspect out of their analyses if they did think about it. Still, since this is a criterion that many students had difficulties with both in the beginning (when they had access to the rubric) as well as in the end of the program, it might be a criterion that is actually too difficult for novices. Visualizing different scenarios with the situations as starting points may actually require experiences that the students have yet to gain.

This is in contrast to the other criterion that many students had difficulties with both at the beginning and the end of the program, i.e. to support the actions they suggested by making reference to course literature or other relevant sources, which is not a criterion likely to depend on students' experiences in the field. Instead, justifying your actions as a teacher on for instance research literature is probably a skill most easily gained during the teacher-education program, as compared to gaining such a skill through working as an in-service teacher.

## Conclusions and Implications

As has been argued in this article, an important part of teacher competence is analyzing complex classroom situations, including understanding students' incentives for behaving in particular ways, and suggesting appropriate action that follows from the analysis. To develop such competencies takes time, however, and it has been argued that teacher education can merely pave the way, and that the students have to learn from their own experiences (Sabers et al., 1991). Still, there are studies showing that teacher education does make a difference. For instance, Blömeke et al. (2008) found significant improvements along several dimensions of teacher competence, when comparing students' performance at the beginning and at the end of teacher education with students from four countries (Germany, South Korea, Taiwan, and USA). These authors report that criteria such as making judgments about lesson goals, and the use of professional terminology, are especially strong indicators of teacher-education effects. The current study also provides some tentative evidence of the development of analytical skills, since the students performed at least at the same level during their last semester as during their first, but without the assistance of a rubric. This interpretation, however, is very optimistic and a more credible conclusion is that there has not been any substantial development during the teacher-education program regarding the analytical skills investigated in this study. This conclusion is further corroborated by the fact that there are weaknesses in students' performances which are consistent throughout the teacher-education program. Similar tendencies can be found in the study by Blömeke et al., where assessment procedures for German students remained at the same low level at the end of teacher education as it was in the beginning. Studies like these can thus give indications of areas that might need to be strengthened in teacher education, acknowledging of course that neither study claims to have a representative sample and therefore that these findings may not necessarily generalize to a larger population of students.

To conclude, this study has shown that the analytical skills of pre-service teachers do not develop substantially during teacher education, since they perform at the same level during the last semester as they did during their first. The only exceptions are changes in two of the criteria assessed: (1) The students have *increased* their awareness about the need to have more information about the students and the context before being able to make well-grounded decisions; (2) The students consider different motives for acting in particular ways *to a lesser extent* during the last semester. There are also criteria in relation to which the students perform consistently low. One of these is the skill of justifying your analysis and suggested actions from course literature or other relevant sources, and since this is a skill that is not likely to require extensive experiences from the field, it might be considered especially problematic that the students have not developed this skill during the teacher-education program. The second criterion where the students perform consistently low, on the other hand, requires them to envisage potential consequences of the situations; something that might actually call for students to meet and handle a range of different situations.

The main implication of this study is that further attention needs to be directed towards pre-service teachers' development of analytical skills. In relation to research, this means that additional research is needed to establish to what extent the learning of such skills can be facilitated during teacher education. This would perhaps be most interesting in relation to "discussing conceivable consequences of different situations," a skill that turned out to be a consistent weakness for the students and a skill that might

be difficult to address. While the "Interactive examination" presents situations for the students to analyze, they do not see the consequences of the actions they suggest. If these actions could somehow be simulated, however, the students could see what consequences their actions had, and they could perhaps also go back and try other solutions to the same problem. In this way students could potentially by aided in their development to discuss conceivable consequences of different classroom situations – by experiencing them vicariously and reflecting on different solutions to the same problem. Another possible area for future research, as suggested by this study, would be to investigate the development of these skills by in-service teachers. For instance: To what extent do teachers develop analytical skills and which factors are important in order for teachers to excel in this area?

The major implication for teacher education would also be to pay closer attention to students' development of analytical skills. While there are studies (e.g. Jonsson, 2010; Metcalf et al., 1996) suggesting that the learning of analytical and reflective skills can be improved by for instance reflecting on simulated situations and by clarifying expectations, this study indicates that not all students develop such skills during their teacher education and are still novices at the time they graduate. From an even broader perspective, the results point towards the need for a more integrated curriculum in teacher education, where skills not directly related to subject matter are addressed during the campus-based parts of teacher education and skills not procedural in nature are addressed during the practicum, in order to aid the students in basing their actions on relevant knowledge – i.e. acting knowledgeably.

## References


Berliner, D. C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*, 5-13.

Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society, 24*, 200-212.

Björklund, L-E. (2008). *Från novis till expert: Förtrogenhetskunskap i kognitiv och didaktisk belysning* [From Novice to Expert: Intuition in a Cognitive and Educational Perspective]. Doctoral dissertation, Linköping University, Sweden.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7-74.

Blömeke, S., Paine, L., Houang, R. T., Hsieh, F-J., Schmidt, W. H., Tatto, M. T., Bankov, K., et al. (2008). Future teachers' competence to plan a lesson: first results of a six-country study on the efficiency of teacher education. *ZDM Mathematics Education, 40*, 749–762.

Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 39*, 25-31.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: the power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell.

Elliott, J. (1991). A model of professionalism and its implications for teacher education. *British Educational Research Journal, 17*, 309-318.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*, 27-32.

Hegender, H. (2010). Mellan akademi och profession. Hur lärarkunskap formuleras och bedöms i verksamhetsförlagd lärarutbildning [Between academy and

profession. How teacher knowledge is formulated and assessed in school-based teacher education]. Doctoral dissertation, Linköping University, Sweden.

Jonsson, A. (2010). The use of transparency in the "Interactive examination" for student teachers. *Assessment in Education: Principles, Policy & Practice, 17,* 185-199.

Jonsson, A., Baartman, L. K. J., & Lennung, S. A. (2009). Estimating the quality of performance assessments: The case of an "Interactive examination" for teacher competency. *Learning Environments Research,* 12, 225-241.

Jonsson, A., Mattheos, N., Svingby, G., & Attström, R. (2007). Dynamic assessment and the "Interactive examination". *Educational Technology & Society, 10*, 17-27.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Jönsson, A. (2008). *Educational assessment for/of teacher competency*. Doctoral dissertation, Malmö University, Sweden.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lin, S. S. J. (1999, April). *Looking for the prototype of teaching expertise: An initial attempt in Taiwan*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA, USA.

Mattheos, N., Nattestad, A., Falk Nilsson, E., & Attström, R. (2004). The interactive examination: Assessing students' self-assessment ability. *Medical Education, 38*, 378-389.

Metcalf, K. K., Ronen Hammer, M. A., & Kahlich, P. A. (1996). Alternatives to field-based experiences: The comparative effects of on-campus laboratories. *Teaching and Teacher Education, 12*, 271-283.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal, 28*, 63-88.

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about new modes of assessment in higher education: a review. *Assessment & Evaluation in Higher Education, 30*, 325–341.

Taconis, R., Van der Plas, P., & Van der Sanden, J. (2004). The development of professional competencies by educational assistants in school-based teacher education. *European Journal of Teacher Education, 27*, 215-240.

van den Berg, E. (2001). An exploration of the use of multimedia cases as a reflective tool in teacher education. *Research in Science Education, 31*, 245-265.

## About the Author

**Anders Jonsson** holds a position as Assistant Professor at Kristianstad University College, Sweden, and is also a researcher in Educational Research at the Centre for Profession Studies (CPS) at Malmö University, Sweden. His research interest is in assessment, especially the assessment of professional competency in higher education, but is also involved in projects concerning assessment of science in compulsory school.

**Sven A. Lennung** is an Associate Professor in Educational Sciences. He has recently returned to research after several years as head of a publishing firm specializing in course literature for higher education.

Correspondence concerning this article should be addressed to Anders Jonsson, Malmö University, Centre for Profession Studies, SE-205 05 MALMÖ, Sweden.
E-mail: anders.jonsson@mah.se
Phone: 0046-701-751668

The research presented here was supported by Malmö University Centre for Profession Studies, CPS.

Educational Measurement and Evaluation Review (2011), Vol. 2, 17-34


http://pemea.club.officelive.com/ EMEReview.aspx

# Temperament Styles of Children from Samoa and the United States

**Carmelo M. Callueng**
*University of Florida*

**Desmond M. Lee Hang**
*National University of Samoa*

**Richard DLC. Gonzales**
*University of Santo Tomas, Philippines*

**Ainslie Chu Ling-So'o**
*Ministry of Education, Sports and Culture, Samoa*

**Thomas D. Oakland**
*University of Macau*

**Abstract** Age, gender, and cross-national differences among children ages 9 through 16 in Samoa and the United States are examined on four bipolar temperament styles: extroversion-introversion, practical-imaginative, thinking-feeling, and organized-flexible. Samoan children generally prefer extroverted to introverted, practical to imaginative, thinking to feeling, and organized to flexible styles. Gender differences are found in practical-imaginative styles. Compared to males, relatively more females prefer a practical style. Age differences are found on extroversion-introversion and thinking-feeling styles. While Samoan children in all age groups generally prefer an extroverted style, an increased preference for an introverted style is seen at each older age. On the other hand, younger children display a somewhat balanced preference for thinking and feeling styles while older children generally prefer a thinking style. Cross-national differences are found in extroversion-introversion, practical-imaginative, and organized-flexible styles. In contrast to children in the U.S., those in Samoa are more likely to prefer extroverted, practical, and organized styles.

**Keywords:** *temperament, Student Styles Questionnaire, children, Samoan children*

## Introduction

Temperament is generally understood as stylistic and relatively stable traits that subsume intrinsic tendencies to act and react in somewhat predictable ways to people, events, and stimuli (Teglasi, 1998a; 1998b). Temperament traits generally are characterized as predispositions to display behaviors, with no assurance that people, events, and stimuli always will elicit the same temperament behaviors. Temperament traits appear early in life (e.g., Goldsmith, et al., 1987; Thomas & Chess 1977) and thus are assumed to have a biological origin, one tempered both by one's environment as

well as personal choice (Bates & Wachs, 1994; Goldsmith, et al., 1987; Kagan, 1994b; Keogh, 2003; Oakland, Glutting, & Horton, 1996). Age and gender also are assumed to influence temperament.

Children's temperament can have a substantial impact on their behaviors, including their personal motivation, learning styles, peer and family relationships, and values (Bates & Wachs, 1994; Hofstede, 1980; Joyce, 2010; Keirsey & Bates, 1984; Keogh, 2003; Lawrence, 1982; Oakland, et al., 1996). For example, temperament impacts vocational interests in children as young as 8 (Oakland, et al, 2001) and may help distinguish children who display conduct and oppositional defiant disorders (Joyce & Oakland, 2005). Style preferences of sighted and non-sighted children also were compared (Oakland, Banner, & Livingston, 2000). Thus, knowledge of children's temperament shows promise for use in understanding the impact of temperament on children's behaviors.

## Temperament Styles Theory and Current Assessment

Jung's temperament theory (1953, 1971) helped launched considerable research and test development (Bassett & Oakland, 2009). Jung (1921, 1959) attributed individual differences to inborn, possibly genetic or physiological qualities mediated by one's environment. He emphasized the importance of two attitudes (i.e. extraversion-introversion) together with four mental functions (i.e. thinking-feeling and sensation-intuition) that impact the apprehension of stimuli. His writings focused heavily on extraversion-introversion, given his belief that they define important individual differences. However, for Jung, temperament is understood best by examining interactions between extroversion-introversion and the four mental functions (i.e. thinking feeling and sensation-intuition), and not by focusing on each dyadic pair separately.

Briggs and Myers' successful application of Jung's theory in test form, the Myers-Briggs Type Indicator (MBTI; Myers and McCaulley, 1985), brought Jung's theory to life and set the stage for its dissemination and practical applications, with the MBTI reportedly one of the most widely used measures in the world (Myers et al., 1998). In developing the MBTI, Briggs and Myers utilized Jung's extroversion-introversion, separated his thinking-feeling and sensation-intuition into two separate traits, and added a fourth: judging-perceptive.

Oakland and his co-authors operationalized Briggs and Myers' theory of temperament in their Student Styles Questionnaire for children and youth (SSQ; Oakland et al., 1996). The SSQ is based on the premise that temperament results from an interaction between biologically coded qualities, environmental qualities and personal choice. The SSQ assesses four temperament style dimensions for ages 8-17 (Table 1): extroversion-introversion, practical-imaginative (consistent with the MBTI's sensing-intuitive), thinking-feeling and organized-flexible (consistent with the MBTI's judging-perceiving).

Table 1
*Descriptions of Temperament Qualities (from Horton & Oakland, 1997)*

**Extroversion-Introversion Styles**

This dimension describes individuals' orientations to the outer world of people and events around them. Those with extroverted preferences generally are energized by contact with people, while those with introverted preferences generally derive energy from their inner world of thoughts.

*Those with an extroverted style generally learn by:*
- *talking*
- *enjoy large groups*
- *have many interests & friends*
- *respond quickly*

*Those with an introverted style generally learn by:*
- *reflecting and writing*
- *prefer small groups or solitude*
- *have a few interests and close friends*
- *respond with hesitance & caution*

**Practical-Imaginative Styles**

This dimension describes individuals' orientations to ideas and experience. Those with practical preferences generally attend to facts and objects, while those with imaginative preferences generally view the world in terms of possibilities and insights.

*Those with a practical style generally are:*
- *realistic/pragmatic*
- *understand things*
- *literally enjoy sequential learning*
- *notice details*

*Those with an imaginative style generally are:*
- *insightful/visionary/theory oriented*
- *enjoy metaphor/symbolism*
- *learn by insight/intuitive leaps*
- *notice themes/generalizations*

**Thinking-Feeling Styles**

This dimension describes individuals' orientations for making decisions. Those with thinking preferences generally use objective standards to make decisions and strive for fairness, while those with feeling preferences generally use personal standards to make decisions and strive for harmony.

*Those with thinking style generally are:*
- *analytical/quizzical*
- *value logic over sentiment*
- *display brief/businesslike interactions*
- *strive for fairness/truth/justice*

*Those with feeling style generally are:*
- *trusting/sympathetic/seek harmony*
- *value sentiment over logic*
- *tactful/friendly interactions*
- *strive for harmony compassion*

Cont. Table 1

| Organized-Flexible Styles | |
|---|---|
| This dimension describes individuals' orientations as to when they make decisions. Those with organized preference styles generally prefer to finalize decisions and have issues settled as soon as possible while those with flexible preference styles generally prefer to delay decisions and keep their options open. | |
| *Those with organized style generally* | *Those with flexible style generally are:* |
| • *want to plan/schedule* | • *flexible in commitments* |
| • *persist, are dependable* | • *seek opportunity for play* |
| • *keep personal space neat* | • *tolerate disorder of possessions* |
| • *enjoy predictable/structure* | • *enjoy surprise/adaptive to change* |

## Temperament Style Preferences among U.S. Children

The U.S.-based New York longitudinal study reported that temperament differences between males and females appear shortly after infancy and increase with age on the following New York longitudinal study qualities: adaptability, approach/withdrawal, activity, and sensory threshold (Chess & Thomas, 1991). During the period from 4 months to 4 years, males are more adaptable and approaching than females. Between ages 8 to 12, males display higher levels of activity and sensitivity (Maziade, et al., 1986).

Older children and youth in the United States generally prefer extroverted, imaginative, and organized styles. They display age related differences on extroversion-introversion styles (i.e., a preference for extroversion increases from 8 to 13), on practical-imaginative styles (i.e., a preference for an imaginative style generally increases with age), and on organized-flexible styles (i.e., a preference for a flexible style generally increases with age) (Oakland, et al., 1996; Bassett, 2005).

Gender differences also exist (Oakland, et al., 1996; Bassett & Oakland, 2009). More females than males prefer feeling and organized styles while more males than females prefer thinking and flexible styles. Gender differences on thinking-feeling appear early, at least by age 8, are sustained through adulthood, and may be somewhat universal (Myers & McCaulley, 1985; Hammer & Mitchell, 1996; Myers, et al., 1998).

## Cross-national Studies on Children's Temperaments

Research by Oakland and his colleagues is cognizant of emic and etic approaches (Berry, Poortinga, Segall, & Dasen, 1992) in their international studies of children's temperament, including children in Australia (Oakland, Faulkner, & Bassett, 2005), Costa Rica (Oakland & Mata, 2007), Gaza (Oakland, Alghorani, & Lee, 2006), Greece (Oakland & Hatzichristou, 2010), India (Oakland, Singh, Callueng, & Goen, 2011), Japan (Callueng, de Carvalho, Isobe, & Oakland, under review), Hungary (Katona & Oakland, 2000), Nigeria (Oakland, Mogaji, & Dempsey, 2006), Pakistan (Oakland, Rizwan, Aftab, & Callueng, under review), People's Republic of China (Oakland & Lu, 2006), Romania (Oakland, Illiescu, Dinca, & Dempsey, 2009), South Africa (Oakland & Pretorisus, 2009), South Korea (Oakland & Lee, 2010), United

States (Bassett & Oakland, 2009), Taiwan (Oakland et. al., under review), Venezuela (Leon et al, 2009), and Zimbabwe (Oakland, Mpofu, & Sulkowski, 2007).

Emic approaches examine culture-specific traits while etic approaches examine whether traits and behaviors are universal and independent of one or more cultures. Initial research on any trait typically is directed toward describing the trait. Thus, the general purposes of this cross-national research are first to examine commonly displayed temperament traits of children within a country or region and then compare them with children in other countries or regions. This strategy is consistent with cross-national studies by McCrae & Costa (1997) and others (e.g., Berry, Poortinga, Segall, & Dasen, 1992; Macdaid, McCaulley, & Kainz, 1991; Plomin & Dunn, 1986) that examine the possible universality of temperament and personality traits.

### Purposes of the Research

The purpose of this research is to describe temperament style preferences in a sample of Samoan children at four age groups, examine possible gender and age differences among them, and compare their temperament style preferences with children in the U.S. A discussion of temperament styles preferences among U.S. children is not a primary focus of this study and can be found elsewhere (Oakland et al., 1996; Bassett, 2005; Bassett and Oakland, 2009). Data on U.S. children are included to provide a direct cross-national comparison.

The following questions are addressed in this study: 1) Do Samoan children display differences in their preferences for extroversion-introversion, practical-imaginative, thinking-feeling, or organized-flexible styles? 2) Do they display gender and age differences on these temperament styles? 3) Do Samoan and U.S. children differ in their preferences for these styles?

## Method

### Participants

Data were collected on 400 Samoan public school children who reside in the nation's capital, Apia, a city with approximately 38,000 residents (Government of Samoa Department of Statistics, 2006). Sample sizes were 100 at each of the following four age groups: 9-10, 11-12, 13-14, and 15-16, with 50 males and 50 females in each age group. Classes were selected randomly, and all children within the classes completed the Student Styles Questionnaire (SSQ: Oakland et. al., 1996). The school serves children from low and middle to upper lower class families who dwell in urban Apia and rural villages.

A sample of 800 U.S. children was drawn from the U.S. Student Style Questionnaire's standardization sample (Oakland et al., 1996), with 100 males and 100 females in each of the following age groups: 9-10, 11-12, 13-14, and 15-16. The U.S. standardization data were designed to reflect 1990 U.S. Bureau of the Census data. Thus, the US sample is stratified on five variables: age, gender, race/ethnicity, geographic region, and school type. The subject pool of 7,902 public and private school children ranged in ages 8 through 17 years. Three racial/ethnic groups (Anglo-Americans, African-Americans, and Hispanics) were represented proportionately; approximately 50% were males in each racial-ethnic group.

## Instrument

The Student Styles Questionnaire (SSQ; Oakland, Glutting, & Horton, 1996) is patterned after the Jungian constructs popularized by the Myers-Briggs Type Indicator (Myers & McCaulley, 1985). The SSQ, a self-report paper and pencil group administered measure of temperament type for children ages 8 through 17, is completed within approximately 20 minutes. Each of its 69 forced-choice items has two alternatives that provide for an assessment of preferred behaviors associated with one of four bipolar traits: extroversion (E) or introversion (I), practical (P) or imaginative (M), thinking (T) or feeling (F), and organized (O) or flexible (L). An example of an item assessing extroversion-introversion follows: After school, I most prefer to a) spend time with others, 2) spend time alone. The EI scale has 23 items, the PM scale has 16 items, the TF scale has 10 items, and the OL scale has 26 items. Additionally, 6 items provide information simultaneously on two scales.

Test-retest reliability coefficients, derived over an 8 month period, are .80, .67, .70, and .78 for EI, PM, TF, and OL respectively. Results of factor analyses studies indicate the SSQ's factor structure is consistent and stable for U.S. children who differ by age, gender, and racial-ethnic group (Stafford & Oakland 1996a; 1996b). Factor analytic studies of data from children from seven countries generally found a stable factor structure and thus support the use of the SSQ internationally (Benson, Oakland, & Shermis, 2009). External validity, using contrasted groups, convergent validity, and divergent validity, provides additional strong support for the SSQ's validity (Oakland et al., 1996).

## Procedures

The SSQ was reviewed for use with Samoan children. A review of the SSQ's 69 items and directions by the second and fourth authors found them to be generally suitable for Samoan children and consistent with their culture. In accord with the guidelines in test adaptation (Hambleton, Merenda, & Spielberger, 2005), the U.S. SSQ items and directions were first translated into Samoan language by the fourth author and retranslated into English language by the second author. Both translators were very proficient in Samoan and English languages and had prior experience in test translation. The entire test adaptation process was supervised by the third author who has a doctoral degree in research and evaluation and has done substantial work related to test development and adaptation. The translated test was administered to Samoan children consistent with the administration procedures described on the SSQ record form.

## Data Analysis

Temperament typically is considered to be a type rather than a continuous quality (Bassett, 2005; Bassett & Oakland, 2009; Buss & Plomin, 1984; Hall & Lindzey, 1978; Jung, 1946; Lawrence, 1982; Macdaid, et al., 1991; Plomin, & Dunn, 1986; Rothbart & Jones, 1998; Teglasi, 1998b; Thomas & Chess, 1977). Personality also can be and often is viewed in its type form (McCrae & Costa, 1997). This belief guided the data interpretation methods.

The frequency of Samoan and U.S. children expressing a preference for each of the eight types were determined in the following way. Individual responses on each

of the 69 items were examined to determine whether a child selected more options from one of the two bipolar types. For example, among the 23 extroversion-introversion items, children who selected more extroverted than introverted options were classified as extroverted. Conversely, children who selected more introverted options were classified as introverted. Children who selected an equal number of options on a scale (e.g., extroversion-introversion) display no discernable preference on that bipolar type and thus were dropped from subsequent analyses on that scale. Less than three percent of the sample was excluded from any one of the four bipolar traits due to their experiencing equal number of item preferences on one of the four scales.

Data were analyzed using frequencies and are reported using percentiles to promote understanding. Tests for significance of a proportion and chi-square ($\chi^2$) analyses were used to test whether the frequency of Samoan children who prefer either extroversion or introversion, practical or imaginative, thinking or feeling, and organized or flexible styles differs significantly by age, gender, or compared to preferences of children in the U.S. Z-scores equal to or greater than 1.96 are considered to be significant. Possible differences between Samoan and U.S. children in reference to the total group, for males, females, and four age groups are examined through chi-square analyses. A .05 significance level was set for all analyses. Thus, p-values equal to or less than .05 indicate groups differ statistically on a temperament dimension.

## Results

### Temperament Styles of Children from Samoa

**Preferences for Extroverted and Introverted Styles**. More Samoan children prefer an extroverted (70%) than an introverted (30%) style ($\chi^2$ (1) = 64.00, $p < .001$). For these and other results refer to Table 2. Gender differences are not significant ($\chi^2$ (1) = .05, $p > .05$). Age differences are evident between 9-10 and 11-12 ($\chi^2$ (1) = 10.70, $p < .01$), 9-10 and 13-14 ($\chi^2$ (1) = 15.79, $p < .001$), and 9-10 and 15-16 ($\chi^2$(1) = 21.59, $p < .001$). Compared to older children, more 9-10 year-old children display a preference for an extroverted style.

**Preferences for Practical and Imaginative Styles.** More Samoan children prefer a practical (89%) than an imaginative (11%) style ($\chi^2$ (1) = 224.71, $p < .001$). Gender differences are significant ($\chi^2$ (1) = 5.92, $p < .05$). Relatively more females than males prefer a practical style. Age differences are not evident.

**Preferences for Thinking and Feeling Styles**. More Samoan children prefer a thinking (63%) than a feeling (37%) style ($\chi^2$(1) = 23.29, $p < .001$). Gender differences are not significant ($\chi^2$ (1) = .22, n. s.). Age differences are evident for 9-10 and 11-12 ($\chi^2$ (1) = 4.69, $p < .05$), 9-10 and 13-14 ($\chi^2$ (1) = 7.95, $p < .01$), and 9-10 and 15-16 ($\chi^2$ (1) = 7.95, $p < .01$). Older children are more inclined to prefer a thinking style while younger children display a somewhat balanced preference for thinking and feeling styles.

**Preferences for Organized and Flexible Styles**. More Samoan children prefer an organized (98%) than a flexible (2%) style ($\chi^2$ (1) = 369.49, $p < .001$). Gender differences are not significant ($\chi^2$ (1) = 1.85, $p > .05$). Age differences are not evident in all groups: 9-

10 and 11-12 ($\chi^2$ (1) = 2.99, $p > .05$), 9-10 and 13-14 ($\chi^2$ (1) = .12, $p > .05$), 9-10 and 15-16 ($\chi^2$ (1) = .34, $p > .05$), 11-12 and 13-14 ($\chi^2$ (1) = 1.89, $p > .05$), 11-12 and 15-16 ($\chi^2$ (1) = .21, $p > .05$), and 13-14 and 15-16 ($\chi^2$ (1) = .00, $p > .05$).

Table 2
*Temperament Preferences for Children from Samoa and the United States for Total Group, Gender, and Four Age Groups (by percent)*

| Group | E | I | P | M | T | F | O | L |
|---|---|---|---|---|---|---|---|---|
| *Samoa* | | | | | | | | |
| Age 9-10 | 88 | 12 | 92 | 8 | 48 | 52 | 99 | 1 |
| Age 11-12 | 69 | 31 | 83 | 17 | 65 | 35 | 97 | 3 |
| Age 13-14 | 64 | 36 | 90 | 10 | 70 | 30 | 98 | 2 |
| Age 15-16 | 59 | 41 | 92 | 8 | 70 | 30 | 98 | 2 |
| Male | 70 | 30 | 85 | 15 | 62 | 38 | 96 | 4 |
| Female | 69 | 31 | 93 | 7 | 65 | 35 | 99 | 1 |
| Total | 70 | 30 | 89 | 11 | 63 | 37 | 98 | 2 |
| *United States* | | | | | | | | |
| Age 9-10 | 46 | 54 | 41 | 59 | 55 | 45 | 84 | 16 |
| Age 11-12 | 53 | 47 | 41 | 59 | 53 | 47 | 77 | 23 |
| Age 13-14 | 62 | 38 | 41 | 59 | 50 | 50 | 65 | 35 |
| Age 15-16 | 55 | 45 | 47 | 53 | 50 | 50 | 56 | 44 |
| Male | 55 | 45 | 43 | 57 | 72 | 28 | 65 | 35 |
| Female | 55 | 45 | 39 | 61 | 33 | 67 | 79 | 21 |
| Total | 55 | 45 | 42 | 58 | 52 | 48 | 71 | 29 |

*Note.* E = extroverted; I = introverted; P = practical; M = Imaginative; F = feeling; O = organized; L = Flexible.

## Cross-national Differences on Temperament Styles of Children from Samoa and U.S.

**Preferences for Extroverted and Introverted Styles**. Samoan and U.S. children differ in their preferences for extroverted and introverted styles ($\chi^2$ (1) = 4.80, $p < .05$). Compared to U.S. children, relatively more Samoan children prefer an extroverted style. These differences are significant for males ($\chi^2$ (1) = 4.80, $p < .05$) and females ($\chi^2$ (1) = 4.16, $p < .05$). Compared to U.S. children, relatively more Samoan males and females prefer an extroverted style. Age differences are evident for 9-10 year-old ($\chi^2$ (1) = 39.89, $p < .001$) and 11-12 year old ($\chi^2$ (1) = 5.38, $p < .05$) children. More 9-10-year-old Samoan children display a preference for an extroverted style while their U.S. peers are more inclined to prefer an introverted style. In the 11-12 year-old group, relatively more Samoan children prefer an extroverted style compared to their U.S. peers.

**Preferences for Practical and Imaginative Styles**. Samoan and U.S. children differ in their preferences for practical and imaginative styles ($\chi^2$ (1) = 48.88, $p < .001$). More Samoan children display a preference for a practical style while more U.S. children display preference for an imaginative style. These differences are significant for males ($\chi^2$ (1) = 32.28, $p < .001$) and for females ($\chi^2$ (1) = 64.97, $p < .001$). Age differences are evident for all age groups: 9-10 ($\chi^2$ (1) = 58.38, $p < .001$), 11-12 ($\chi^2$ (1) = 37.44, $p < .001$), 13-14 ($\chi^2$ (1) = 53.12, $p < .001$), and 15-16 ($\chi^2$ (1) = 47.76, $p < .001$).

Compared to their U.S. peers who are more inclined to prefer an imaginative style, Samoan children in all age groups are more inclined to prefer a practical style.

**Preferences for Thinking and Feeling Styles**. Samoan and U.S. children do not differ in their preferences for thinking and feeling styles ($\chi^2$ (1) = 2.48, n. s.). Gender differences are significant for females ($\chi^2$ (1) = 20.49, $p < .001$) and not for males ($\chi^2$(1) = 2.26, n. s.). More Samoan females prefer a thinking style while more U.S. females prefer a feeling style. On the other hand, males from both countries display a preference for a thinking style. Age differences are evident for 13-14 ($\chi^2$ (1) = 8.33, $p <$ .01) and 15-16 ($\chi^2$ (1) = 8.33, $p <$ .01). Samoan children in both age groups display a preference for a thinking style while their U.S. peers display a balanced preference for thinking and feeling styles.

**Preferences for Organized and Flexible Styles**. Samoan and U.S. children differ in their preferences for organized and flexible styles ($\chi^2$ (1) = 27.83, $p < .001$). Compared to U.S. children, relatively more Samoan children prefer an organized style. These differences are significant for males ($\chi^2$ (1) = 30.61, $p < .001$) and for females ($\chi^2$ (1) = 20.67, $p < .001$). Compared to U.S. males and females, more Samoan males and females prefer an organized style. Age differences are evident for all groups: 9-10 ($\chi^2$ (1) = 14.47, $p < .001$), 11-12 ($\chi^2$ (1) = 17.68, $p < .001$), 13-14 ($\chi^2$ (1) = 36.11, $p < .001$), and 15-16 ($\chi^2$ (1) = 49.80, $p < .001$). Compared to their U.S. peers, more Samoan children prefer an organized style.

## Discussion

### Preference for Extroverted and Introverted Styles

Samoan children show a decided preference for an extroverted style, with 70% preferring this style and 30% preferring an introverted style. An extroverted style is preferred by both males and females and across ages. The greater preference for extroverted style in Samoan children reflects their natural propensity to be with peers and seek peer approval. Children are regarded as normal and well functioning when they put forth their energy to actively explore the environment, searching for both physical and social adventures (Matas, Arend, & Sroufe, 1978). This dynamic interaction of children in their environment enables them to acquire a variety of social skills, leading to a wider range of interpersonal relationships, to learn new behaviors, and to be open to new experiences (Higgins & Parsons, 1983). The increasing social network of extroverted children allows them to shift and mingle comfortably with family members and their peers as well (Brown, Mounts, Lamborn, & Steinberg, 1993).

Preference for an extroverted style also reflects the Samoan cultural idea of relational self or *va* (Tamasese, Peteru, & Waldegrade, 2005). *Va* is regarded as a cultural value that fosters loyalty and love for family ties among Samoans. In addition, customs and traditions in Samoan culture emphasize communal practices and thus promote positive interactions and socialization in Samoan children. In classroom setting, extroverted Samoan children may work well in a group setting and benefit more when exposed to cooperative learning methods (Rzoska & Ward, 1991).

Although Samoan children express a greater preference for an extroverted style, an increased preference for an introverted style is seen in older children. Children who prefer an introverted style generally derive their energy from themselves and are described to learn best by having time to think about and reflect upon what they have learned. Cultures also generally expect older children to be more self-directed and independent, qualities associated with introversion In Samoan culture, this characteristic is exemplified in the practice of *le-tautala* or student silence wherein older children are more inclined to practice *le-tautala* and think through the implications of their actions and experiences (Lee-Hang, 2011).

Differences between Samoan and U.S. children are apparent on extroversion-introversion styles. Compared to U.S. children, Samoan children are more likely to prefer an extroverted style. Young Samoan children display a preference for an extroverted style and relatively more of their U.S. peers are inclined to prefer an introverted style. The tendency of children to prefer extroversion to introversion also has been found among children in Australia (Oakland, Faulkner, & Bassett, 2005), Costa Rica (Oakland & Mata, 2007), Greece (Oakland & Hatzichristou, 2010), India (Oakland, Singh, Callueng, & Goen, 2011), Japan (Callueng, de Carvalho Filho, Isobe, & Oakland, under review), Pakistan (Oakland, Rizwan, Aftab, & Callueng, under review), People's Republic of China (Oakland & Lu, 2006), , Romania (Oakland, Illiescu, Dinca, & Dempsey, 2009), South Africa (Oakland & Pretorisus, 2009), Venezuela (Leon et al, 2009), and Zimbabwe (Oakland, Mpofu, & Sulkowski, 2007).

## Preferences for Practical and Imaginative Styles

Samoan children display a remarkably high preference for a practical style, with 89% endorsing this style and only 11% endorsing imaginative style. Gender differences are evident. Compared to males, more Samoan females prefer a practical style.

Practical-imaginative styles relate to life orientation and processing of ideas and experiences. Thus, a practical style preference is likely to be linked to learning style preferences vis-à-vis gender differences. For example, females are likely to prefer learning strategies that involve repetitive drill and practice and that focus mainly on details and task that require knowledge-level learning (Severiens & Ten Dan, 1994). These learning orientation and practices are descriptive of the practical style that Samoan female children are more inclined to display. On the other hand, males are more inclined to learn through logical analysis and active questioning, to engage in deep processing, and to use evidence when formulating conclusions. These learning preferences and behaviors are related to imaginative style that Samoan male children are more inclined to.

Differences between Samoan and U.S. children are apparent on practical-imaginative styles, with Samoan children more likely to prefer a practical style and U.S. children more likely to prefer an imaginative style. These differences are consistent in both males and females. A practical style preference also is displayed by children from Gaza, Greece, Hungary, Japan, Nigeria, Pakistan, People's Republic of China, Philippines, Romania, South Africa, Venezuela, and Zimbabwe.

## Preferences for Thinking and Feeling Styles

Samoan children show a general preference for a thinking style, with 63% endorsing this style and 37% endorsing a feeling style. The preference for a thinking style is consistent in both older male and female children. On the other hand, younger children display a somewhat balanced preference for thinking and feeling styles.

Similarities among male and female Samoan children on thinking-feeling styles were not expected as this finding is inconsistent with results from other studies that generally report that males prefer a thinking style and females prefer a feeling style (Joyce, 2010; Myers, et. al., 1998; Hammer & Mitchell, 1996; Oakland, et. al, 1996; & Myers & MacCaulley, 1985).

While there is a somewhat consistent and established pattern of gender differences in thinking-feeling styles in children cross-culturally, the socialization of emotions in Samoan children may explain their strong preference for a thinking style among both females and males. Samoan culture does not teach children to express their feelings; emotions usually are reserved and difficult to express. Samoans use the term *"ma"* to suppress their true feelings. Therefore, children are encouraged not to talk about their feelings, especially in the presence of adults (Pereira, 2005). Such cultural practice may contribute to making the seemingly inherently biologically-based gender differences in thinking-feeling styles more latent among Samoan children.

When examined cross-nationally, differences in a preference for thinking style is relatively greater in magnitude among Samoan children (63%) than U.S. children (52%). Moreover, a thinking style is more likely to be preferred by Samoan (65%) than U.S. (33%) females. Age differences are evident. Older Samoan children show a greater preference for a thinking style while their U.S. peers show a greater preference for a feeling style. Children from Australia, People's Republic of China, and Romania also reported a general preference for a thinking style.

## Preferences for Organized and Flexible Style

Samoan children display a decided and remarkably high preference for an organized style, with almost (98%) all children preferring this style. Moreover, a preference for an organized style is highly consistent among males and females and remains stable and high from ages 9 through 16. However, the persistent high preference for organized style even among older children is contrary to the findings that preference for an organized style generally decreases as age increases (Joyce, 2010).

A typical parenting style in Samoan culture involves discipline and order- an indication of a greater proclivity for an organized style. Samoan children's orientation to an organized style begins during children's formative years by the family and reinforced through disciplined regimen in formal learning and socialization in school and in the village. Samoan children are exposed to some organized structure in doing home chores at a very early age. They typically enjoy being with their families, accepting domestic duties, and want and expect to do them correctly, to the delight of adults (Ochs & Izquierdo, 2009). In addition, communal activities of the other institutions in the village (e.g., church) require Samoan children and adolescents to follow orders from adults, to display strong adherence to rules, and to complete tasks well and in a timely manner.

Both Samoan and U.S. children generally prefer an organized style. However, this preference is more prominent in Samoan children. In U.S. children, males are more likely than females to prefer a flexible style; additionally, this preference increases with age (Oakland et. al., 1996; Bassett, 2005). However, in Samoan children, both males and females show high preference for an organized style and remain stable from ages 9-10 through ages 15-16. Children in most countries display a general preference for an organized style: Australia, Costa Rica, Gaza, Greece, Hungary, India, Japan, Nigeria, Pakistan, People's Republic of China, Romania, South Africa, Venezuela, and Zimbabwe. Only children from South Korea display a general preference for flexible style.

## Implications for Assessment of Samoan Children's Temperament Styles

This study examined age, gender, and cross-national differences on temperament styles of children from Samoa and the U.S. The SSQ was used to obtain baseline data on temperament qualities of Samoan children. Although the SSQ has been popularly utilized in research and educational program evaluation studies, it was primarily developed and made available in the U.S. to describe children's temperament and thus help to explain their behaviors. Items of the SSQ were carefully selected to characterize preferences of children rather than categorize or label them (Oakland, Glutting, & Horton, 1996). Because of the impact of temperament styles on children's behavior (Horton & Oakland, 1977; Keirsey & Bates, 1984; Keogh, 2003; Lawrence, 1982; Oakland, Glutting, & Horton, 1996), uses of SSQ are diverse and can be described into several broad categories: identifying talent, adjusting for possible weaknesses, enhancing personal and social development, promoting an understanding of others, assessing learning styles, promoting educational development, and exploring prevocational interests (Oakland, Glutting, & Horton, 1996).

The use of SSQ internationally has been promising and supported by a stable factor structure (Benson, Oakland & Shermis, 2009). To date, several countries (e.g., Romania, South Korea, and Taiwan) have made successful adaptations of the SSQ and currently are used to assess children.

There is no measure of temperament preferences for Samoan children ages 7-18 years. The SSQ Samoan version shows promise in use with this population. However, its psychometric properties should be further examined and normative data should be established. An understanding of a child's temperament preferences may be enhanced by the use of three methods to interpret SSQ scores (Oakland, Glutting, & Horton, 1996).

The first method, referred to as the basic level of interpreting the SSQ, considers eight basic styles grouped into four bipolar scales: extroverted-introverted; practical-imaginative; thinking-feeling; and organized-flexible. Furthermore, the four bipolar scales can provide insights onto the problem solving process in terms of a) whether a child acquires energy (e.g., either from the outer world of people, things and events and thus is extroverted or from the inner world of thoughts and ideas and thus is introverted, b) whether a child attends to either facts (i.e., is practical) or possibilities and imaginative solutions (i.e., is imaginative), c) how a child makes decisions (e.g., either based on logic and objective standards and thus is thinking or on personal standards derived from beliefs and attitudes and thus is feeling), and d) whether a child

makes a decision as early as possible and thus is organized or delays them and thus is flexible.

The second method, patterned from the book of Keirsey and Bates, *Please Understand Me: Character Temperament Type* (1984), interprets temperament preferences by combining two styles. The combinations of practical-organized, practical-flexible, imaginative-thinking, and imaginative-feeling styles can provide valuable information associated with social and family relationships, learning preferences and applications, career preferences, and personal values.

The third method of SSQ interpretation that utilizes the combinations of four preferred styles was introduced by Myers and McCaulley (1985) and Myers (1987) and advanced by the Association for Psychological Types. The resulting 16 style combinations (e.g., extroverted, practical, thinking, and organized) stress the importance of understanding the interactions of the styles and thus enhances a meaningful interpretation of the child's SSQ profile.

## Limitations and Future Research

We believe this is the first study that examines temperament styles among Samoan children. The availability of current data from other temperament studies of Samoan children would facilitate an examination of the reliability of these findings, including gender and age trends. Moreover, Samoan children living in the Apia city does not represent the Samoan children generally. Thus, further research that examines Samoan children's temperament styles should employ a larger sample and rigorous sampling method that resembles the actual social and demographic characteristics of children and youth. An evaluation of the psychometric properties of the SSQ-Samoa is needed, including its factor structure and other validity data. Additional research on temperament styles of children in special groups (e.g., children with learning disabilities, social-emotional problems, behavioral difficulties, gifted children) would facilitate this evaluation. Information on relationships between temperament styles and emotional-behavioral, career, values, and other relevant variables would contribute to this effort and may support the general adaptation of SSQ- Samoa.

## References

Bassett, K. (2005). Nature, nurture, and temperament: Comparisons of temperament styles displayed by U.S. students. Doctoral dissertation, University of Florida. *Dissertation Abstracts International, 122(101/A),* 8993.

Bassett, K. and Oakland, T. (2009). Temperament preferences for children ages 8 through 17 in a nationally represented sample. In J. Kaufman (Ed). *Intelligent testing: Integrating psychological theory and clinical practice* (pp. 30-52). Boston: Cambridge University Press.

Bates, J. E., & Wachs, T. D. (Eds.), (1994) *Temperament: Individual differences of the interface of biology and behavior*. Washington DC: American Psychological Association.

Benson, N., Oakland, T., & Shermis, N. (2009) Cross-national invariance of children's temperament. *Journal of Psychoeducational Assessment, 27*, 3-16.

Berry, J. W. Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (1992). *Cross-cultural psychology: Research and applications.* Cambridge: Cambridge University Press.

Brown, B. B., Mounts, N., Lamborn, S. D., & Steinberg, L. (1993). Parenting practice and peer group affiliation. *Child Development, 64,* 467-482.

Buss, A. H., & Plomin, R. (1984). *Temperament: Early developing personality traits.* Hillsdale, NJ: Erlbaum.

Callueng, C., de Carvalho Filho, M. K., Isobe, M., & Oakland, T. (under review). Temperament Styles of children from Japan and the United States.

Department of Statistics (2006). 2006 Census of Population & Housing. Apia: Department of Statistics, Government of Samoa.

Goldsmith, H. H., Buss, A. H., Plomin, R., Rothbart, M. K., Thomas, A., Chess, S., Hinde, R. A., & McCall, R. B. (1987). Roundtable: What is temperament? Four approaches. *Child Development, 58,* 505-529.

Hambleton, R. K., Spielberger, C., & Merenda, P. (eds) (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* Hillsdale, NJ: Lawrence-Erlbaum. Hall, C.S., & Lindzey, G. (1978). *Theories of personality* (3rd ed.). New York: Wiley.

Hall, J., & Altmaier, E. (Eds.) (2008). *Global promise: Quality assurance and accountability in profesional psychology.* New York: Oxford UniversityPress.

Higgings, E. T. & Parsons, J. (1983). Social cognition and the social life of thechild: Stages as subcultures. In E.T. Higgins, D. Ruble, & W.W. Hartup (Eds.), *Social cognition and social behavior: development perspectives* (pp. 15-62). New York: Cambridge University Press.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills, CA: Sage.

Horton, C., & Oakland, T. (1997). *Student Styles Questionnaire classroom application booklet.* San Antonio, TX: The Psychological Corporation.

Joyce, D. (2010). *Essentials of temperament assessment.* New York: Wiley.

Joyce, D., & Oakland, T. (2005). Temperament differences among children with conduct disorder and oppositional defiant disorder, *The California School Psychologist, 10,* 125-136.

Jung, C. G. (1946). *Psychological types* (H.G. Baynes, Trans.). New York: Harcourt (Original work published 1921).

Jung, C. G. (1953). *Two essays on analytical psychology.* (R.F.C. Hull, Trans.). New York: Pantheon Books (Original work published 1943).

Jung, C. G. (1959). The archetypes and the collective unconscious. In *Collected works. Vol. 9, Part I.* Princeton University Press (Original work published 1936-1955).

Jung, C. G. (1971). *Psychological types.* (R. F. C. Hull, Revision of Trans. By H. G. Baynes). Princeton, NJ: Princeton University Press (Original work published 1921).

Kagan, J. (1998). Biology and the child. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 105-176). New York: Wiley.

Kagan, J. (1989). Temperamental contributions to social behavior. *American Psychologist, 44,* 668-674.

Kagan, J. (1994a). *Galen's prophecy: Temperament in human nature.* New York: Basic Books.

Kagan, J. (1994b). Inhibited and uninhibited temperaments. In W.B. Carey & S.C. McDevitt, (Eds.), *Prevention and early intervention: Individual differences as risk factors for the mental health of children* (pp. 35-41). New York: Brunner.

Matas, L., Arend, R., & Sroufe, L.A. (1978). Continuity of adaptation in the second year: The relationship between quality of attachment and later competence. *Child Development*, *49*, 547-556.

Mazel.Katona, N., & Oakland, T. (2000). The development of temperament in Hungarian children. *Hungarian Journal of Psychology, 1*, 17-29.

Keirsey, D., & Bates, M. (1984). *Please understand me*. Del Mar, CA: Prometheus Nemesis Book Company.

Keogh, B. (2003). *Temperament in the classroom*. New York: Brooks Publishing.

Lawrence, G. (1982). *People types and tiger stripes: A practical guide to learning styles* (2nd ed.). Gainesville, Florida: Center for Applications of Psychological Type.

Lee-Hang, D. M. (2011). Fa'afatāmanu talafeagai mo lesona Fa'asaienisi: O le tu'ualalo mo aoga a faiaoga saienisi fa'aḍiḍi. A culturally appropriate formative assessment in science lessons: Implications for initial science teacher education. Unpublished doctoral thesis. University of Waikato: Hamilton, New Zealand.

Leon, C., Oakland, T., Wei, Y., & Berrios, M. (2009) Venezuelan children's temperament styles and comparison with their United States peers. *Revista Interamericana de Psicología/Interamerican Journal of Psychology, 43,* 407-415.

Macdaid, G., McCaulley, M., & Kainz, R. (1991). *Atlas of type tables*. Gainesville, FL: Center for the Application of Psychological Types.

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52(5)*, 509-516.

Myers, I. B. (1987). Introduction to type (4th ed.). Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Oakland, T., Glutting, J., & Horton, C. (1996*). Student Styles Questionnaire.* San Antonio, TX: The Psychological Corporation.

Oakland, T., Stafford, M., Horton, C., & Glutting, J. (2001). Temperament and vocational preferences: age, gender, and racial-ethnic comparisons, *Journal of Career Assessment, 9,* 297-314.

Oakland, T., Banner, D., & Livingston, R. (2000) Temperament-based learning styles of visually impaired children. *Journal of Visual Impairment and Blindness, 94*, 26-33.

Oakland, T., Joyce, D., Horton, C., & Glutting J. (2000). Temperament-based learning styles of male and female gifted and non-gifted children. *Gifted Child Quarterly,* 44, 183-189.

Oakland, T. Faulkner, M., & Bassett, K. (2005). Temperament styles of children from Australia and the United States, *Australian Educational and Developmental Psychologist*, *19*, 35- 51.

Oakland, T., Alghorani, M. A., & Lee. D. H. (2006). Temperament-based learning styles of Palestinian and US children. *School Psychology International*, *28*, 110-128.

Oakland, T., Mogaji, A., & Dempsey, J. (2006). Temperament Styles of Nigerian and U.S. Children. *Journal of Psychology in Africa*, *16*, 27-34.

Oakland, T., & Lu, L. (2006). Temperament styles of children from the People's Republic of China and the United States. *School Psychology International, 27*, 192-208.

Oakland, T., & Mata, A. (2007) Temperament styles of children from Costa Rica and the United States. *Journal of Psychological Type, 67,* 91-102.

Oakland, T., Mpofu, E., & Sulkowski, M. (2007). Temperament styles of Zimbabwe and U.S. children. *Canadian Journal of School Psychology, 21*, 139-153.

Oakland, T. Illiescu, D., Dinca, M., & Dempsey, A. (2009). Temperament styles of Romanian children. *Psihologia Sociala, 22*, 70-84.

Oakland, T., & Lee, D. H. (2010) Temperament styles of children from South Korea and the United States. *School Psychology International, 31*, 77-94.

Oakland, T., & Hatzichristou, S, (2010) Temperament styles of children from Greece and the United States. *School Psychology International, 31,* 422-437.

Oakland, T., Rizwan, M., Aftad, S., & Callueng, C. (under review). Temperament styles of children from Pakistan and the United States.

Oakland, T., & Pretorius, J. (2009). Temperament styles of children from South Africa and the United States. *School Psychology International 29*, 627-639.

Oakland, T., Singh, K., Callueng, C., & Goen, A. (2011). Temperament styles of Indian and U.S. children. *School Psychology International*. Advance online publication. doi: 10.1177/0143034311403041

Ochs, E., & Izquierdo, C. (2009). Responsibility in childhood: Three developmental trajectories. *Ethos: Journal of the Society for Psychological Anthropology 37*(4), 391-413.

Oxford, R., & Nuby, J. (1998). Learning style preferences of Native American and African American secondary students. *Journal of Psychological Type, 44,* 5-19.

Pereira, J.A. (2005). Aspects of primary education in Samoa: Exploring student, parent and teacher perspectives. Unpublished doctoral thesis. University of Otago: Dunedin, New Zealand.

Plomin, R., & Dunn, J. (1986). *The study of temperament: Changes, continuities, and challenges*. Hillsdale, NJ: Erlbaum.

Rothbart, M. K., & Jones, L. B. (1998). Temperament, self-regulation, and education. *School Psychology Review, 27,* 479-491.

Rzoska, K. M., & Ward, C. (1991). The effects of cooperative and competitive learning: Methods on the mathematics achievement, attitudes toward school, self-concepts, and friendship choices of Maori, Paheka, and Samoan children. *New Zealand Journal of Psychology, 20*, 17-24.

Severiens, S. E., & Ten Dan, G. T. M. (1994). Gender differences in learning styles: A narrative review and quantitative analysis. *Higher Education, 27,* 487-501.

Stafford, M., & Oakland, T. (1996a). Validity of temperament constructs using the Student Styles Questionnaire: Comparisons for three racial-ethnic groups. *Journal of Psychoeducational Assessment, 14,* 109-120.

Stafford, M., & Oakland, T. (1996b). Racial-ethnic comparisons of temperament constructs for three age groups using the Student Styles Questionnaire. *Measurement and Evaluation in Counseling and Development, 19* (2), 100-110.

Strelau, J., & Angleitner, A. (1994). Cross-cultural studies on temperament: Theoretical considerations and empirical studies based on the pavlovian temperament survey. *Personality and Individual Differences, 16* (2), 331-342.

Tamasese, K., Peteru, C., & Waldegrave, C. (2005). Ole taeao afua, the new morning: A qualitative investigation into Samoan perspectives on mental health and culturally appropriate services. *Australian and New Zealand Journal of Psychiatry, 39*(4), 300-309.

Teglasi, H. (1998a). Introduction to the mini-series: Implications of temperament for the practice of school psychology. *School Psychology Review, 27*, 475-478.

Teglasi, H. (1998b). Temperament constructs and measures. *School Psychology Review, 27,* 564-585.

Thomas, A., & Chess, S. (1977). *Temperament and development*. New York: Brunner/Mazel.

Wedding, D., & Stevens, M. (Eds.) (2004). *The handbook of international psychology*. New York: Routledge.

## About the Authors

**Carmelo M. Callueng** is a doctoral candidate in school psychology at the University of Florida. His research interests include test development and use, children's temperaments, adaptive behaviors, and cognitive-behavior therapy. He received the 2009 American Psychological Association Division 33 (Intellectual and Developmental Disabilities) award for outstanding research.

**Richard DLC Gonzales** is a Professorial Lecturer at the Graduate School of University of Santo Tomas, Manila, Philippines where he teaches assessment, tests and measurement, psychometrics and statistics. He completed his doctorate, major in research and evaluation and cognate in psychology from University of the Philippines-Diliman. His research interests include motivation in foreign language learning, test development, adaptation and use, classroom assessment practices, beliefs and preferences, teaching-learning and school-based assessment.

**Desmond M. Lee-Hang** is a science lecturer at the National University of Samoa. He recently completed his doctorate in education from the University of Waikato. His research interests include culturally-appropriate formative assessments, Samoan language issues in science education, and the influence of culture on science learning, teaching and assessment.

**Ainslie Chu Ling-So'o** is the Language Consultant Specialist in the Curriculum, Materials and Assessment Division of the Ministry of Education, Sports and Culture in Samoa. She completed her Master of Arts in Linguistics at the Australian National University in 1997. She has been a primary school teacher for many years but has now devoted herself in developing materials in the Samoan language to support the Ministry's bilingual language policy. She also is the co-author of a language book entitled *Say it in Samoan*.

**Thomas D. Oakland** is a Professor Emeritus at the University of Florida and Visiting Professor of Psychology at the University of Macau, the University of Hong Kong, and Beijing University-Zhuhai. He received APA's 2003 Award for Distinguished Contributions to the Advancement of Psychology Internationally. He serves as presidents of the International Foundation for Children's Education and the International Association of Applied Psychologists' Division of Psychological Assessment and Evaluation. He is past-presidents of the International School Psychology Association and the International Test Commission. He has worked in more than 45 countries. Dr. Oakland has authored or edited 14 books, 200 chapters, another 200 papers, and authored or assisted in the development of 23 psychological tests.

Educational Measurement and Evaluation Review (2011), Vol. 2, 35-61


http://pemea.club.officelive.com/ EMEReview.aspx

# Monitoring Teacher Trainees' Mathematical Competence in an Accelerated Teacher Education Program

Karoline Afamasaga-Fuata'i
*National University of Samoa*

**Abstract** The paper focuses on the longitudinal impact the first mathematics content (MC) course had on teacher trainees' mathematical performance, during an Accelerated Diploma in Education Program (ADEP), as repeatedly measured by a mathematics diagnostic test based on content typically taught at primary and early secondary levels. With the first diagnostic test (MDT1) administered after the first teaching block, the second diagnostic test (MDT2) was administered prior to the second teaching block 12 months later with the third mathematics diagnostic test (MDT3) administered at the completion of the MC course's second teaching block. Data from the three diagnostic tests were analysed using the Rasch Measurement Model to calibrate teacher trainees' ability measures over time. Implications for teaching mathematics in the classrooms and preparation of competent teachers of primary mathematics are provided.

**Keywords:** *mathematics competence, ability measures, primary mathematics, effect sizes*

## Introduction

Teacher assistants, already working as volunteers in Samoan primary classrooms, undertook an Accelerated Diploma in Education Program (ADEP), so that they could be certified as competent primary teachers. Of importance was the need to ensure that these teacher assistants graduated with an acceptable level of mathematics competence to adequately prepare them to flexibly and confidently teach the new primary mathematics curriculum. The latter emphasises the promotion of small group activities, discovery-oriented mathematical investigation and exploration and deeper attention to working mathematically through mathematical reflection, communication and modelling/representation, all with the emphasis on valuing students' own thinking, questioning, applying strategies and reasoning. Teacher trainees' content knowledge of the primary mathematics curriculum was objectively diagnosed longitudinally, to ensure satisfactory achievement of an appropriate level of mathematics competence (more than the level they are expected to teach) within the time constraints of the accelerated program.

## Research Purpose and Focus Questions

The longitudinal research monitored the nature of the teacher trainees' (TTs) developing mathematical competence as they undertook the two required mathematics content (MC) and pedagogical (MP) courses over three teaching blocks of the ADEP. The empirical, longitudinal data generated and its grounded findings provided, firstly, *on-going diagnostic information* to identify specific content areas and skills that still require further remediation and development during the course of the ADEP, and secondly, *benchmark evidence of TTs' mathematics achievement* at various key points of the ADEP program. This paper focuses only on the presentation and analyses of the data generated by the first three diagnostic tests in the first two teaching blocks, which enabled the calibration of TTs' ability estimates early in the program.

The overall study's main research question was: *In what ways have TTs' mathematics competence evolved over the duration of their ADEP program?* From this main question, the specific research questions relevant to the data presented here are as follows:

1. What are TTs' base line mathematical abilities and identified areas of difficulties early in their ADEP program?
2. What are TTs' mathematical abilities, areas of improvement and areas of on-going difficulties by the end of the first two teaching blocks?

## Theoretical Framework

The mathematical abilities of the teacher trainees as measured by a mathematics diagnostic test are important and relevant to their current and future job as teachers of primary mathematics. In accordance with Shulman's (1986) theory of teacher knowledge for teaching, there are three categories of subject matter knowledge for teaching (SMKT), namely content knowledge (CoK), pedagogical content knowledge (PCK) and curriculum knowledge (CuK). Content knowledge includes both facts and concepts in a domain and also how knowledge is validated, produced and structured in the discipline while PCK includes both the knowledge of the subject matter and knowledge of the subject matter for teaching. More particularly, *pedagogical content knowledge* is the label used to describe what (among other things) teachers know about which mathematical topics typically cause students difficulty, the nature of the difficulties they have, and how particular examples or explanations can be useful in teaching particular mathematical concepts (Shulman, 1986). Since its identification, researchers have found that PCK plays important roles in teachers' practices and the learning opportunities that such practices create for students. For example, teachers' knowledge of the different strategies that their students use to approach problems is positively correlated with student achievement (Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996). Ball (1990), based on analyses of classrooms, distinguished further teacher's content knowledge for teaching knowledge about mathematics (knowledge of concepts, ideas and procedures and how they work) and knowledge about "doing" mathematics (how one decides that a claim is true, a solution is complete or a representation is accurate). In general, these perspectives theorise that teacher effects on student achievement are driven by teachers' ability to understand and use subject matter knowledge to carry out the tasks of teaching. Empirically identifying the effects of teachers' knowledge on student learning and the kinds of teacher knowledge that matter most in

producing student learning led Hill, Schilling, & Ball (2004) to develop an instrument to measure teachers' mathematical knowledge for teaching elementary school mathematics. The instrument not only captured the actual content teachers taught – e.g., decimals, area measurement – but also the specialised knowledge of mathematics needed for the work of teaching such as knowing how to represent ¼ in diagrams or how to appraise multiple solutions for 35 x 25. This instrument was subsequently used in an empirical study to determine the effects of teachers' mathematical knowledge for teaching on student achievement.

"Mathematics knowledge for teaching" (MKT) has been proposed as a specific form of mathematical knowledge used to carry out the *work of teaching*, which includes explaining terms and concepts to students, interpreting students' statements and solutions, using representations accurately in the classroom, and providing students with examples of mathematical concepts, algorithms, or proofs (Ball & Bass, 2000; Hill, Rowan, & Ball, 2005). Hill et al. found teachers' mathematical knowledge plays a significant role even in the teaching of very elementary mathematics content. These findings further inform that teachers' content knowledge should be at least content-specific and even better specific to the knowledge used in teaching children. Recently, Hill, Ball, & Schilling (2008) shifted their use of MKT by referring to such knowledge as *specialized content knowledge* (SCK), a uniquely pedagogical subset of subject matter knowledge. SCK was defined as "the mathematical knowledge that allows teachers to engage in particular *teaching* tasks" such as following students' mathematical thinking, evaluating the validity of student-generated strategies, and making sense of a range of student-generated solution paths (Hill et al., 2008, p. 377). A case study of a mathematician by Speer & Wagner (2009) found that the mathematician struggled to effectively provide in the moment scaffolding of whole-class discussions toward intended mathematical goals and that pedagogical skills and knowledge that served the mathematician well in traditional lecture classrooms were not sufficient for developing mathematical concepts as they arose from students' incipient and informal ideas.

In this paper, it is argued that for SCK to work effectively in practice, primary teachers need to, first of all, be competent with the content of the curriculum that they plan to teach before they are in a position to productively scaffold the development of their students' thinking and reasoning. Unlike the studies by Hill et al. (2005) and Speer and Wagner (2009) which focused on MKT for certified teachers and mathematicians who are, already teaching, the study reported here focused instead on assessing Samoan teacher trainees' content-specific knowledge and ability to solve items based on this content as part of their teacher education program. Test items covered primary content areas such as whole numbers, fractions, decimals, percentage, operations, multi-digit subtraction, rate, ratio, proportion, area and perimeter, which comprised a significant portion of primary mathematics and including items that a first year secondary student is capable of solving (geometry, probability, and basic algebra). Subsequently, TTs' *mathematical competence* is conceptualised, in this study, as the ability to solve problems based on the relevant syllabus's content, requiring teachers to *have* the appropriate mathematical knowledge and understanding of the specific content areas being examined and *ability to link* mathematics to experiences and *to ask* questions about the application of particular mathematical knowledge (Hogan, 2000). Being mathematically competent with Samoa's Ministry of Education, Sports & Culture (SMESC) Primary & Early Secondary Mathematics (PESM) curriculum is conceptualized as one of several factors influencing teachers' goals and the

ways they work to accomplish those goals as they plan for, reflect on, and enact teaching in Samoan classrooms.

## Literature Review of the Samoan Educational Context

### The New Samoan Primary Mathematics Curriculum

To support the design, development and provision of learning activities based on the new Samoan primary mathematics curriculum, primary teachers are expected to lead whole-class and small-group discussions of mathematical ideas (Ball, 1993; Lampert, 2001; NCTM, 1989, 1991, 2000), thereby promoting mathematical thinking and reasoning and consideration/evaluation of each others' ideas leading towards the development of, reflection on, and communication about, important mathematical concepts. To provide this type of teaching, teachers must take the mathematical contributions that students generate and decide, in the moment, which contributed ideas should be pursued, how to pursue them, which examples should be used, or how even incorrect mathematical ideas can be used to advance the achievement of students' learning outcomes (Speer & Wagner, 2009). As Speer and Wagner pointed out, a challenge in leading whole-class discussions is negotiating what is often seen as a tension between the need to encourage and value students' ideas and the simultaneous use of those ideas to keep the discussion moving in a mathematically productive direction (Ball, 1993; Nathan & Knuth, 2003; Sherin, 2002; Stein, Engle, Smith, & Hughes, 2008; Williams & Baxter, 1996). Findings from Speer and Wagner's (2009) case study imply that, for those with weak content knowledge it will be even more difficult to provide the right level of analytic scaffolding necessary to support students' development of their own mathematical reasoning and justification skills. Therefore, it is important for the TTs in this study to demonstrate competence with the content of the primary mathematics curriculum as part of their preparation to become certified primary teachers.

### National Numeracy Results and Diagnostic Test Results

National results of literacy and numeracy tests, as measured by the Samoa's Primary English and Literacy Levels (SPELL) tests at Year 4 (SPELL I) and Year 6 (SPELL II),[1] at early and middle primary indicate that for each national Year 4 cohort, the percentage of students who do not achieve their national benchmarks in literacy and numeracy (i.e., at-risk), often increases by the time they are assessed again by SPELL II in Year 6, despite completing another year of schooling since SPELL I (Year 4) (Afamasaga-Fuatai, Meyer & Falo, 2010, 2008, 2007). Additional research conducted by Afamasaga-Fuata'i and her colleagues (2010, 2008, 2007), using a diagnostic mathematics test consisting of items mainly based on SMESC's PESM content, further confirmed that the majority of Year 10 secondary students, from four local secondary schools that participated in the studies, also found solving word problems difficult. Using the same diagnostic test, empirical evidence from responses of post-secondary students enrolled in the Foundation program at the National University of Samoa (NUS) verified the existence of difficulties and errors associated with solving word problems, albeit to a relatively lesser extent in comparison to similar findings with the early secondary Year 10 students. Invariantly and

[1] Samoa's Years 4 and 6 are equivalent to Australian NSW Years 3 and 5.

collectively demonstrated by these empirical studies and national results were students' difficulties solving word problems on primary content such as whole numbers, fractions, decimals, operations and measurement, as earlier identified by SPELL II tests, again identified at early secondary at Year 10 and apparently persisting up to the end of secondary level many years later (Afamasaga-Fuata'i, Meyer & Falo, 2010, 2008, 2007).

Collectively, between the continuing negative trends of Years 4 and 6 numeracy national results and empirical findings from Afamasaga-Fuata'i *et al*'s secondary and foundation studies, there is sufficient empirical evidence highlighting an urgent need to innovatively and creatively reform the teaching of mathematics, at both primary and secondary levels, and to align current practices, with more recent developments of educational learning theories such as those of the constructivist (Piaget, 1972) and socio-cultural perspectives (Vygotsky, 1978). Whilst reforms in Samoan schools will need to involve all stakeholders including SMESC and teachers, the research reported here focused on seeking empirical evidence to determine whether the participants of the ADEP, with limited entry mathematics background, exit the ADEP program with an acceptable level of mathematics competence (comparable to those of the graduates from the normal 2-year long DEP), to better prepare new teachers to cope with the challenges of SMESC's new primary mathematics curriculum.

## Methodology

Since the overall research study focused specifically on the longitudinal impact of the first and second mathematics content/pedagogical (MC and MP) courses taught within the ADEP program, on a group of TTs' mathematical performance, the research used an instrumental case study strategy (Punch, 2009) in order to develop an in-depth understanding of these TTs' responses (both correct and incorrect) to a diagnostic test to be administered 5 times over three teaching blocks. It is not the intention of the study to generalize, but rather to understand this case in its complexity and its entirety, as well as in its context to gain valuable insights into ways in which participants, with limited mathematics background, within an accelerated teacher education program, can potentially cope with their learning of mathematics within a compressed time period and supported with regular feedback on their performances on the diagnostics test.

Within the case study approach, a mixed methods triangulation design was used in one phase (Punch, 2009) to obtain complementary quantitative data about TTs' mathematical competence and qualitative data about TTs' solutions. In this design, the concurrent, simultaneous collection but separate analyses of the two types of data are then merged at the interpretation-of-results stage (Creswell & Plano Clark (2007) in Punch, 2009). The choice of quantitative method was to enable objective assessment of the TTs' mathematical competence repeatedly through mathematics diagnostics tests that were strategically administered early within the first teaching block and at the beginning and end of subsequent teaching blocks in which the two MC and MP courses were offered.

About 50 participants enrolled in the ADEP program and they formed the case for an in-depth study of their developing mathematics competence. The study's activities were blended into the approved teaching blocks of the ADEP program.

## Data Collection Procedures

The teacher trainees met as a group during intensive teaching blocks usually coinciding with school holidays. Approximately half of the required contact hours (c.h.) (60 c.h.) for the first MC course were offered in the first teaching block in May 2009 after which the first diagnostic test (MDT1) was administered. A 12-month break followed during which the TTs returned to continue being volunteers at their assigned schools. At the beginning of the second teaching block that followed in May 2010, the second diagnostic test (MDT2) was administered prior to the beginning of the one-week intensive teaching block for the completion of the first MC course. The third diagnostic test (MDT3) was administered at the end of the one-week intensive teaching block. The quantitative data included the TTs' correct and incorrect responses while the qualitative data were the TTs' actual solutions to test items from the various tests. The last two diagnostic tests (MDT4 and MDT5) were administered in September 2010. This paper presents data from the first three diagnostic tests only to answer the paper's focus questions.

The 38-item mathematics diagnostic test (MDT) (see Appendix A for brief item descriptions) was previously used in assessing a number of cohorts of pre-service teachers both at the Foundation and DEP programs (Afamasaga-Fuatai, et al., 2010, 2008, 2007). MDT items were based on the content of SMESC's primary and early secondary mathematics curriculum, which is mathematics content that any primary teacher should be competent in. The different versions of the diagnostic tests were designed to have some common items between consecutive tests to facilitate test equating across the 5 diagnostic tests: MDT1 to MDT5 in accordance with the Rasch Model of Measurement. Doing so enabled (and would enable) the linking of ability estimates across MDT1 to MDT5 by anchoring item estimates of common items on previous values to ensure that actual changes in ability estimates between tests were valid and reflected actual changes/improvement in candidate's abilities (Bond & Fox, 2001) with respect to the latent trait (in this study, mathematical competence relative to the content covered by the items).

## Data Analysis

The broad goals of the analysis were to: (a) monitor the progress (or lack of it) made by the TTs during their ADEP, (b) identify areas of improvement (i.e., gains and advancements) and persistent difficulties and then use this information pedagogically to inform subsequent teaching and (c) to provide timely feedback to the TTs soon after the completion of a teaching block. To carry out the analysis, TTs' responses to test items were collected including their actual solutions.

Detailed analyses of TTs' performance at the level of the test as a whole (e.g., Rasch analysis, item analysis and error analysis) helped the researcher/teachers more effectively target topic areas (content, knowledge and skills) to improve in subsequent teaching. With the provision of kidmaps soon after testing, both teachers and students identified areas that needed further work and improvement. Also, item analysis information from Rasch analysis provided useful information on which items tended to have the largest disparities (between success and error percentages or high error percentages), thereby identifying topics most in need of focused intervention and further development.

**Quantitative Data Analysis.** Test responses were marked Correct, Incorrect or Blank and then coded 1, 0 or B respectively, in preparation for analysis. All test results from MDT1 to MDT3 were analysed according to the Rasch Measurement Model (Adams & Khoo, 1996; Rasch, 1980) using the Quest software to provide on-going estimates of TTs' mathematical abilities relative to the content that was examined by the diagnostic test items.

The Rasch Model arises from a fundamental requirement: *that the comparison of two people is independent of which items may be used with the set of items assessing the same variable*. It is considered that the researcher is deliberately developing items that are valid for the purpose and that meet the Rasch requirements (Rasch analysis, 2005). Analysing data according to the Rasch Model (i.e., a Rasch analysis) gives a range of details (e.g., infit mean squares [ims], outfit mean squares [oms] and standardized infit *t* [infit *t*] and outfit *t* [outfit *t*]), which checks whether or not adding the scores is justified in the data. This is called the *test of fit* between the data and the model. It is common practice to accept items as fitting if their infit and outfit mean square values lie within a specified range (often 0.83 to 1.20) around 1.00. Fit statistics are transformed to produce a *t* statistic (mean of zero and a standard deviation of one) and the critical values are usually set at ±2 (see Curtis [2004] and Bond & Fox [2001]). Infit mean squares that are greater than 1.0 (underfit) indicate unmodeled noise or other source of variance in the data while values less than 1.0 (overfit) indicate the model predicts the data too well. Both the magnitude of infit and outfit statistics (and other fit indices) and their corresponding *t* values warrant consideration as do the structure of the samples of items and persons that lead to the data sets that are analysed.

If the data do fit the model adequately for the purpose of the test, then the Rasch analysis also linearises the total score, which is bounded by 0 and the maximum score on the items, into measurements. The linearised value (in logits) is the *location of the person* on the unidimensional continuum – the value is called a *parameter* in the model and there can be only one number in a unidimensional framework. This parameter can then be used in additional analyses more readily than the raw total score, which has floor and ceiling effects (Rasch Analysis, 2005; Bond & Fox, 2001). Unlike the Classical Test Theory (CTT), which simply asserts that the total score is the relevant statistic, with the Rasch model, the total score follows mathematically from the requirement of invariance of comparisons among persons and items (see Rasch analysis, 2005). Fit of the data to the model is therefore paramount and suggests items are working together consistently to define an interpretable construct. The Rasch analysis also provides separation reliability indices to indicate how well the items and persons worked consistently to produce valid measures of the underlying variable. The person estimate reliability is an indication of the precision of the instrument and shows how well individuals can be distinguished by the instrument. Andrich (1982) has shown that this index is virtually identical to the KR-20 or its generalisation, Cronbach alpha. The item estimate reliability shows how well the items that form the scale are discriminated by the sample of respondents. Wright and Masters (1982, pp.90-92) argued that good item separation is a necessary condition for effective measurement.

Rasch analysis results provide evidence from which to develop a nuanced understanding of a students' performance. For example, TTs' ability measures, variable maps and kidmaps provided useful diagnostic and advancement information about students' mathematical knowledge and skills at various key points of the ADEP program.

Variable maps display both item and ability estimates along a common logit continuum; examples are provided later. Kidmaps are one-page graphic displays using a single unidimensional logit scale and 4 sections to display both a student's ability measure and items' difficulty estimates. For example, a kidmap indicates a student's (a) *overall performance* level by positioning his/her ability estimate along the logit continuum located in the middle of the map, (b) *strengths* by displaying the hard and easy items he/she got correct (left-side of continuum) and (c) *weaknesses* by displaying those hard and easy items he/she got incorrect (right-side of continuum). Examples of kidmaps are provided later. The Rasch model predicts that a person has average probability of being successful with items located within his/her ability band (ability estimate ± standard error [S.E.]), more than average probability with items located below the ability band and less than average probability for items above the ability band.

Comparisons between groups of results were done using a standardized mean difference statistic, referred to as *d* (Cohen, 1988). The latter is a scale-free measure of the separation between two group means and it quantifies the practical difference between two sets of results or the size of the effect of a treatment. Calculating *d* for any comparison involves dividing the difference between the two group means by either their average (pooled) standard deviation or by the standard deviation of the control group. This calculation results in a measure of the difference between the two group means expressed in terms of their common standard deviation or that of the untreated population. Thus, a *d* of .25 indicates that one-quarter standard deviation separates the two means (Valentine & Cooper, 2003; Thalheimer & Cook, 2002).

**Linking Ability Measures across the Diagnostic Tests.** Allowing for common items between consecutive diagnostic tests meant Rasch analysis of subsequent test's responses can be anchored on item estimates of the common items from the previous test. Through item anchoring between tests, ability estimates of TTs can be compared across the 3 tests (Bond & Fox, 2001). There were 33 common items between MDT1 and MDT2 with 21 common items between MDT3, MDT2 and MDT1.

**Qualitative Data Analysis.** A cognitive approach to the analysis of TTs' actual solutions enabled the identification of common error types. From the latter, relevant knowledge and skills were ascertained as those most in need of further improvement which then informed the design of focused intervention activities to develop the relevant knowledge and skills in subsequent teaching blocks.

**Triangulation of Findings.** Triangulation from the Rasch analyses of quantitative data and qualitative error analyses enabled the identification of (a) changes in ability estimates, (b) areas of improvement (gains and advancements), and (c) persistent areas of difficulties over time. Collectively (a) to (c) were used to formulate answers to the paper's focus questions.

This paper presents the Rasch analyses results of TTs' responses from the first three diagnostic tests, namely, item and ability estimates, variable maps, some item analysis data and example kidmaps. Error analysis results are not included in this paper.

## Results

### Fit between Data and Rasch Model

Results from the Rasch analysis of responses to the first diagnostic test (n=45), using the Quest software, provided ims, oms, infit *t*, and outfit *t* values which checked whether the data fits the Rasch model. For MDT1, item fit statistics ims (mean = 1.02) and oms (mean = 0.80) values were around one. On the other hand, item infit *t* (mean = 0.12, standard deviation [SD] = 0.95) and outfit *t* (mean = 0.09, SD = 0.63) values were within expected (mean and standard deviation) values, see Table 1. Therefore, both fit statistics (i.e., ims, oms, infit *t* and outfit *t*) suggest reasonable fit of the data to the Rasch model. Only Item 5 on mental multiplication of two decimal fractions (ims = 1.71, oms = 1.80, infit *t* = 2.9 and outfit *t* = 1.6) showed some randomness or underfit indicating more variation in the responses than expected by the Rasch model. However, given the importance of this topic (decimals) in primary mathematics, the item was still retained in the analysis (Bond & Fox, 2001). Relevant fit statistics for MDT2 and MDT3 are also provided in Table 1; these further corroborated the fit of the data to the Rasch Model.

Table 1
*Rasch Results for the First Three Diagnostic Tests*

| | MDT1 | MDT2 | MDT3 | | MDT1 | MDT2 | MDT3 |
|---|---|---|---|---|---|---|---|
| Item Estimates | | | | Case Estimates | | | |
| Mean | 0.00 | 0.00 | 0.32 | Mean | -1.32 | -1.19 | -0.59 |
| SD | 2.10 | 1.66 | 1.77 | SD | 1.39 | 1.08 | 1.29 |
| Reliability | 0.93 | 0.98 | 0.95 | Reliability | 0.86 | 0.82 | 0.88 |
| Item Fit Statistics | | | | Case Fit Statistics | | | |
| Infit MS | | | | Infit MS | | | |
| Mean | 1.02 | 1.20 | 1.08 | Mean | 0.99 | 1.08 | 1.06 |
| SD | 0.22 | 0.78 | 0.37 | SD | 0.28 | 0.27 | 0.22 |
| Outfit MS | | | | Outfit MS | | | |
| Mean | 0.80 | 1.52 | 1.03 | Mean | 0.80 | 1.52 | 1.03 |
| SD | 0.35 | 1.51 | 0.64 | SD | 0.50 | 1.47 | 0.61 |
| Infit t | | | | Infit t | | | |
| Mean | 0.12 | 0.24 | 0.23 | Mean | 0.05 | 0.28 | 0.25 |
| SD | 0.95 | 1.15 | 1.32 | SD | 0.97 | 1.03 | 0.91 |
| Outfit t | | | | Outfit t | | | |
| Mean | 0.09 | 0.39 | 0.24 | Mean | 0.18 | 0.46 | 0.21 |
| SD | 0.63 | 1.13 | 1.12 | SD | 0.60 | 1.04 | 0.81 |
| Items with: | | | | Cases with: | | | |
| Zero scores | 4 | 6 | 3 | Zero scores | 0 | 0 | 0 |
| Perfect scores | 0 | 1 | 0 | Perfect scores | 0 | 0 | 0 |
| Cronbach alpha | 0.85 | 0.83 | 0.87 | | | | |

*Note*. MS – Mean Squares

### Reliability Indices

Two important instrument parameters are reported in Rasch analyses conducted using Quest (Adams & Khoo, 1996) item and person estimate reliability. For MDT1,

reliability indices for item estimates and person estimates were relatively close to one at 0.93 and 0.86 respectively. These indices indicated that the items and persons worked consistently together in the same direction to provide valid estimates. The Cronbach alpha for MDT1 was 0.85. Similarly for MDT2 and MDT3, their item and person estimate reliability indices and Cronbach alpha were high at around 1.00 (see Table 1).

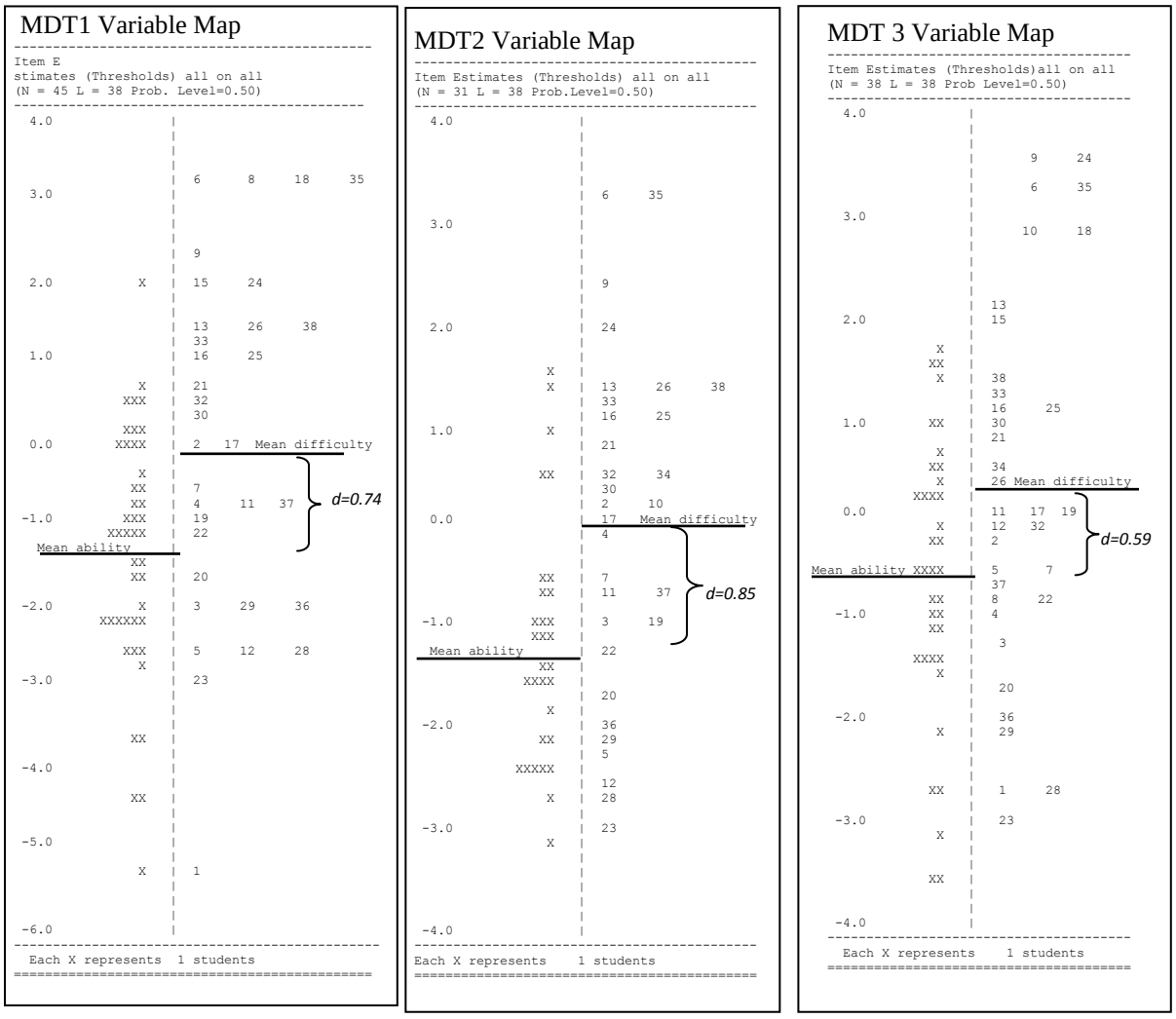
## Item and Ability Estimates

The Rasch analysis of MDT1 data and calibration of ability estimates theoretically sets the mean of the item difficulty estimates at 0.0 logits. The mean of subsequent ability estimates was -1.32 logits (SD = 1.39 logits) (see Table 1). The negative mean ability indicated that the TTs found the test difficult. The MDT1 variable map (Figure 1), showing the distributions of both the ability estimates (as X on the left-side of middle dotted line) and item estimates (as item numbers to the right of middle dotted line) along the common logit continuum, further corroborated that the test was difficult for the TTs.

## Comparison of Item and Person Distributions Within-Tests

An inspection of the MDT1 variable map showed that the means of the person distribution and item distribution do not align. Specifically, item difficulty mean (0.00 logits) was much higher than person ability mean (-1.32 logits). The practical difference between item and person distributions (i.e., inter-distribution gap) was moderate-sized ($d$=0.74) as measured by Cohen's $d$. This moderate effect size indicated that about three-quarters standard deviation separated the two means. Extending this inter-distribution gap analysis within tests for MDT2 and then MDT3 (using the respective values provided in Table 1), Cohen's $d$ increased for MDT2 ($d$ = 0.85), indicating that the practical difference or inter-distribution gap was larger after a 12-month break. However, the inter-distribution gap became moderate sized again for MDT3 ($d$ = 0.59) after the intensive 1-week teaching block. Visually corroborating these within-test inter-distribution gaps and practical differences are the 3 variable maps in Figure 1. Collectively, the effect sizes and variable maps indicated that despite the tests becoming increasingly more difficult (increasing item means), the TTs' mean abilities were also increasing positively in the same direction with the practical difference (or inter-distribution gap) becoming greater in MDT2 after being away on a 12-month break. That the inter-distribution gap narrowed and practical difference reduced to moderate size directly following the second, intensive teaching block demonstrated some gains and advancement with the TTs' learning of mathematics as a result of week-long intervention activities.

Figure 1
*Variable Maps*




## Changes in Item Difficulty Estimates across Tests

For the second diagnostic test (MDT2), with estimates of 25 common items anchored on MDT1 estimates (excluding the 4 MDT1 zero-score items), mean of MDT2 item estimates remained at 0.00 logits but with a narrower spread (SD = 1.66) around the mean (see Table 1 and Figure 1). For MDT3, with estimates of 17 common items anchored on MDT1 estimates (excluding the 4 MDT1 zero-score items), mean item estimate increased from 0.00 to 0.32 logits with a slightly wider spread of estimates around the mean (from 1.66 to 1.77 logits). Apparent was an increase in item difficulties (from MDT1 and MDT2 to MDT3) which could be linked to the presence of 17 new items in MDT3. In fact, whilst the practical difference between MDT1 and MDT2 item estimates in terms of Cohen's *d* was zero, that between MDT2 and MDT3 was small (*d* = 0.19).

## Item Analysis Data for Zero-Score Items Across Tests

Item analysis data are provided in Table 2 for those items that had zero scores (0% success rate, that is, highest disparity between success and error rates) in any one of the three tests as these represent the most difficult items. Also provided are the corresponding item analysis data of the same or variation items in the other two tests for comparison purposes.

In MDT1, the 4 zero-score items included: Item 10 on likely outcomes of tossing a coin, Item 14, the student: professor problem, Item 27 on ratio of width: perimeter given that the rectangle has length that is twice the width and Item 31 about average weight of a crystal. For MDT2, there were 6 zero-score items and 1 perfect-score item (multiplication of two single-digit numbers) whilst there were 3 zero-score items in MDT3. The MDT2 zero-score items were part of the common items between MDT1 and MDT2 and they included: 3 of the MDT1 zero-score items, namely, Items 14, 27 and 31 and new MDT2 zero-score items, Items 8, 15, and 18. The most number of zero-score items occurred in MDT2 after the 12-month break indicating difficulties with retaining previous learning from the first teaching block.

**Item 10: Likely Outcomes of Tossing a Coin (MDT1) -** Item 10 was the fourth MDT1 zero-score item, which in MDT2 had 22.6% (7/31) success rate with reduced incorrect (45.2% [14/31]) and baulked (32.3% [10/31]) rates compared to the corresponding MDT1 data (64.4% [29/45] incorrect and 35.6% [16/45] baulked rates). Of interest are the mean abilities of those that unsuccessfully attempted Item 10 which were more or less the same in both tests (-1.07 logits for MDT1 and -1.20 for MDT2). Similarly for those that baulked (ignored the item as being too difficult and left it blank); mean abilities were more or less the same in both tests (-1.77 for MDT1 and -1.75 for MDT2). In comparison, those that were successful in MDT2, similarly in MDT3, had relatively higher mean abilities of -0.37 and 0.36 logits respectively, as expected.

**Item 8: Arranging Fractions in Ascending Order (MDT2) -** The new zero-score items in MDT2 included Item 8 on arranging $\frac{5}{6}, \frac{2}{3}, \frac{7}{10}, \frac{3}{5}$ in ascending order; only one person was successful (2.2%) in MDT1. Also, the error rate increased in MDT2 to 96.8% (30/31) (91.1% [41/45] in MDT1) with only one student baulking in MDT2 compared to 3 in MDT1. It was interesting to note that with a new variation in MDT3 (i.e., arrange $\frac{5}{6}, \frac{2}{3}, \frac{6}{7}, \frac{9}{10}, \frac{3}{4}$ in ascending order), the success rate became 55.3% (21/38). Most likely this was due in part to the inability of the new variation to discriminate between the most common error of ordering based on the size of the numerator and the correct method using decimals. More particularly, MDT3's success rate was 55.3% (21/38) in contrast to MDT1's and MDT2's zero error rates. Those that attempted unsuccessfully decreased from 96.8% (30/31) in MDT2 to 42.1% (16/38) in MDT3 with baulked rates remaining the same, around 3%. Of interest is the fact that, despite the poor item variation, a higher success rate than 55.3% was not evident; about 44.7% either still got it wrong or baulked in MDT3 (see Table 2). In general, the item analysis data across the three tests suggested that a fair number of TTs continue to experience difficulties ordering fractions.

Table 2
*Item Analysis Data for Zero-Score Items*

| | MDT1 | | | MDT2 | | | MDT3 | | |
|---|---|---|---|---|---|---|---|---|---|
| Categories | Incorrect | Blank | Correct | Incorrect | Blank | Correct | Incorrect | Blank | Correct |
| Item 8 | | | | | | | | | |
| Count | 41 | 3 | 1 | 30 | 1 | 0 | 16 | 1 | 21 |
| Percent (%) | 91.1 | 6.7 | 2.2 | 96.8 | 3.2 | 0 | 42.1 | 2.6 | 55.3 |
| Mean Ability | -1.39 | -0.80 | -0.06 | -1.20 | -0.91 | NA | -0.35 | -1.44 | 0.74 |
| Item 10 | | | | | | | | | |
| Count | 29 | 16 | 0 | 14 | 10 | 7 | 33 | 3 | 2 |
| Percent (%) | 64.4 | 35.6 | 0 | 45.2 | 32.3 | 22.6 | 86.8 | 7.9 | 5.3 |
| Mean Ability | -1.07 | -1.77 | NA | -1.20 | -1.75 | -0.37 | -0.48 | -2.35 | 0.36 |
| Item 14: | | | | | | | | | |
| Count | 17 | 28 | 0 | 17 | 14 | 0 | 25 | 13 | 0 |
| Percent (%) | 37.8 | 62.2 | 0 | 54.8 | 45.2 | 0 | 65.8 | 34.2 | 0 |
| Mean Ability | -0.64 | -1.73 | NA | -1.05 | -1.36 | NA | -0.42 | -0.91 | NA |
| Item 15: | | | | | | | | | |
| Count | 38 | 4 | 3 | 27 | 4 | 0 | 36 | 1 | 1 |
| Percent (%) | 84.4 | 8.9 | 6.7 | 87.1 | 12.9 | 0 | 94.7 | 2.6 | 2.6 |
| Mean Ability | -1.38 | -1.89 | 0.23 | -1.16 | -1.38 | NA | -0.67 | -0.91 | 0.91 |
| Item 18 | | | | | | | | | |
| Count | 37 | 7 | 1 | 26 | 5 | 0 | 30 | 6 | 2 |
| Percent (%) | 82.2 | 15.6 | 2.2 | 83.9 | 16.1 | 0 | 78.9 | 15.8 | 5.3 |
| Mean Ability | -1.28 | -2.00 | 1.97 | -1.17 | -1.30 | NA | -0.51 | -1.21 | 0.18 |
| Item 27 | | | | | | | | | |
| Count | 23 | 22 | 0 | 8 | 13 | 0 | 22 | 16 | 0 |
| Percent(%) | 51.1 | 48.9 | 0 | 58.1 | 41.9 | 0 | 57.9 | 42.1 | 0 |
| Mean Ability | -0.96 | -1.70 | NA | -0.89 | -1.61 | NA | -0.41 | -0.83 | NA |
| Item 31 | | | | | | | | | |
| Count | 24 | 21 | 0 | 20 | 11 | 0 | 25 | 13 | 0 |
| Percent (%) | 54.5 | 45.5 | 0 | 64.5 | 35.5 | 0 | 65.8 | 34.2 | 0 |
| Mean Ability | -1.06 | -1.04 | NA | -0.88 | -1.75 | NA | -0.42 | -1.57 | NA |

**Item 15: Area of Path within Nested Geometric Shapes (MDT2).** This was a common item throughout all 3 tests. No one in MDT2 got it correct with only 6.7% (3/45) and 2.6% (1/38) success rates in MDT1 and MDT3 respectively. Difficulties might lie in identifying the relevant information from amongst those provided in the diagram to determine area of path.

**Item 18: Total Distance Traveled by Elevator (MDT2).** Item 18 was common to both MDT1 and MDT2 and a new variation in MDT3. Whilst only one person out of 48 in MDT1 got Item 18 correct, no one got it correct in MDT2 suggesting that the single TT from MDT1 who got it correct was unsuccessful 12 months later. With a new variation in MDT3 in which the distance apart of the floors was changed from 3 metres to 3.5 metres, two persons got it correct (5.3% [2/38]) with error rate decreased to 79% (30/38) and baulked rate consistently around 16% (6/38).

**Items 14: Mathematical Relationships, Item 27: Ratio of Width to Perimeter and Item 31: Average Weight of Crystal (MDT1, MDT2 and MDT3).** These 3 items, common throughout the 3 tests consistently remained at zero success rate with the majority attempting them unsuccessfully while the rest baulked (see Table 2). Whilst Item 14 was the student: professor problem requiring students to construct a mathematical equation to

represent the relationship between students and professors, Item 27 was a multi-step word problem on perimeter and ratio with a described multiplicative relationship between length and width and Item 31 on operations with decimals to determine an average weight of a crystal given the total number of crystals and total weight. These consistently zero-score items across tests suggested that the TTs continued to experience difficulties solving multi-step problems and operating with decimals. Quite likely these TTs may not have had the exposure and opportunity to solve such problems in their previous mathematics experiences.

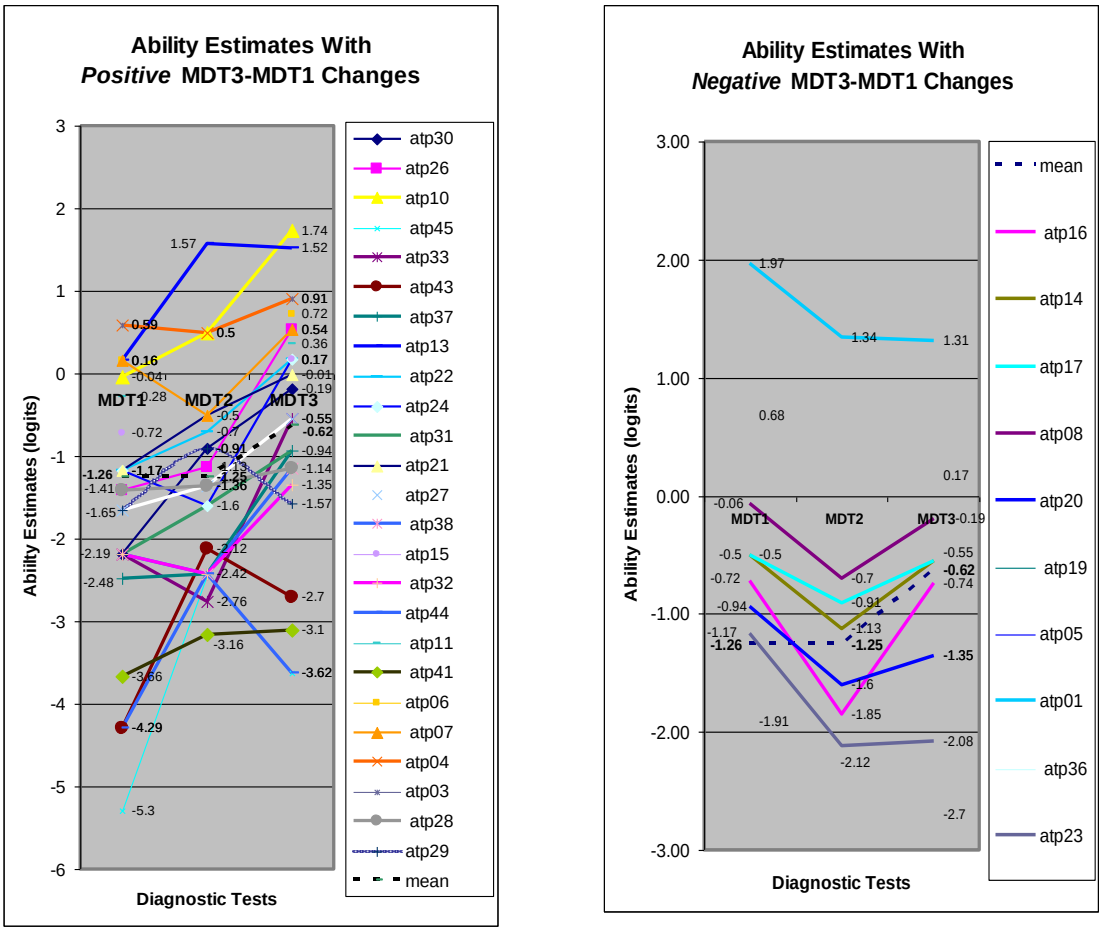
## Changes in Ability Estimates across Tests

The mean ability estimate improved from -1.32 logits in MDT1 to -1.19 logits in MDT2 with a relatively lesser spread of estimates (*SD* = 1.08 compared to 1.39 in MDT1). Using Cohen's *d* (1988), the practical difference between the MDT2 and MDT1 results is small (*d* = 0.10). This is to be expected as the TTs were on teaching practice in the 12 month period since MDT1 and had had no formal mathematics classes during that period. Instead, they were assisting certified teachers in the classrooms. In comparison, MDT3 estimates had an increased mean ability estimate from-1.19 to -0.59 logits (*SD* =1.29). Comparing the MDT3 and MDT2 results, Cohen's *d* (of 0.50) indicated a moderate effect size, suggesting that the one-week intensive teaching block had a moderate effect on the TTs' mathematical competence, comparatively more than the effect size between the first two tests (*d* = 0.10). Of interest also was the moderate effect size (*d* = 0.54) between MDT1 and MDT3 ability estimates. Overall, it appeared that, the effect of the MC's intensive teaching blocks on the TTs' mathematical competence was of moderate size.

**Mathematical Ability Changes between MDT1 and MDT3.** Further in-depth examination of individual teacher trainee's longitudinal ability estimates across the 3 tests showed that some TTs demonstrated positive linear changes from MDT1 to MDT2 and through to MDT3 whilst others were either erratic (decreased and then increased or vice-versa) or demonstrated negative linear changes. When baseline ability estimates at MDT1 were compared to those of MDT3, some ability changes were positive and others negative (see the two graphs in Figure 2). Of interest is the location of the mean ability estimates (dotted line) relative to (a) the positive logit values and (b) those of each teacher trainee. The former indicated that each test's mean ability estimate (-1.26, -1.25, -0.62 logits) was consistently in the negative region and the latter demonstrated whether or not individual performance was above group average or not.

Out of 48 TTs, 52% (25/48) demonstrated positive (MDT3-MDT1) ability changes and are shown in Figure 2a. In contrast, 21% (10/48) had negative MDT3-MDT1 ability changes and are shown in Figure 2b. The rest of the TTs (i.e., 27% [13/48]) undertook either MDT1 or MDT3 but not both and so his/her MDT3-MDT1 ability change could not be determined. Of the 52% with positive MDT3-MDT1 ability changes, 27% (13/48) had MDT3 ability estimates located above-the-group-mean value of -0.62 logits compared to the 25% (12/48) that were located below-the-group-mean estimate. Whilst it was encouraging to note that the effect of the intensive teaching blocks was: (a) for the group, of moderate size (*d* = 0.54), (b) for some TTs, positive (i.e., MDT3 ability estimates were located higher than baseline ability estimates), and (c) for an even fewer TTs, MDT3 ability estimates were above-group-mean and in the greater-than-1.00-logit region (see Figure 2a).

Figure 2
*Longitudinal Ability Estimates*




(a) (b)

## Example Kidmaps

To illustrate the nature of the TTs' developing mathematical competence and type of diagnostic information provided by a kidmap, two examples are presented in Figures 3 and 4. The first example is of Candidate ATP10 who had a positive MDT3-MDT1 ability change and a MDT3 ability estimate greater than 1.00 logits. ATP10's longitudinal ability estimates were -0.04, 0.50 and 1.74 logits for MDT1, MDT2 and MDT3 respectively (see Figure 2a). The second example is of Candidate ATP23 who had a negative MDT3-MDT1 ability change and a MDT3 ability estimate less than -2.00 logits. ATP23's longitudinal ability estimates were -1.17, -2.12 and -2.08 logits for MDT1, MDT2 and MDT3 respectively (see Figure 2b). For each teacher trainee, kidmaps from MDT1 and MDT3 are presented to indicate (a) their overall performance, (b) strengths (hard/easy correct items) and (c) weaknesses (hard/easy incorrect items).

**Mathematical Performance of Candidate ATP10 -** ATP10's overall mathematical performance in MDT1 was calibrated to be -0.04 logits and is located on the logit

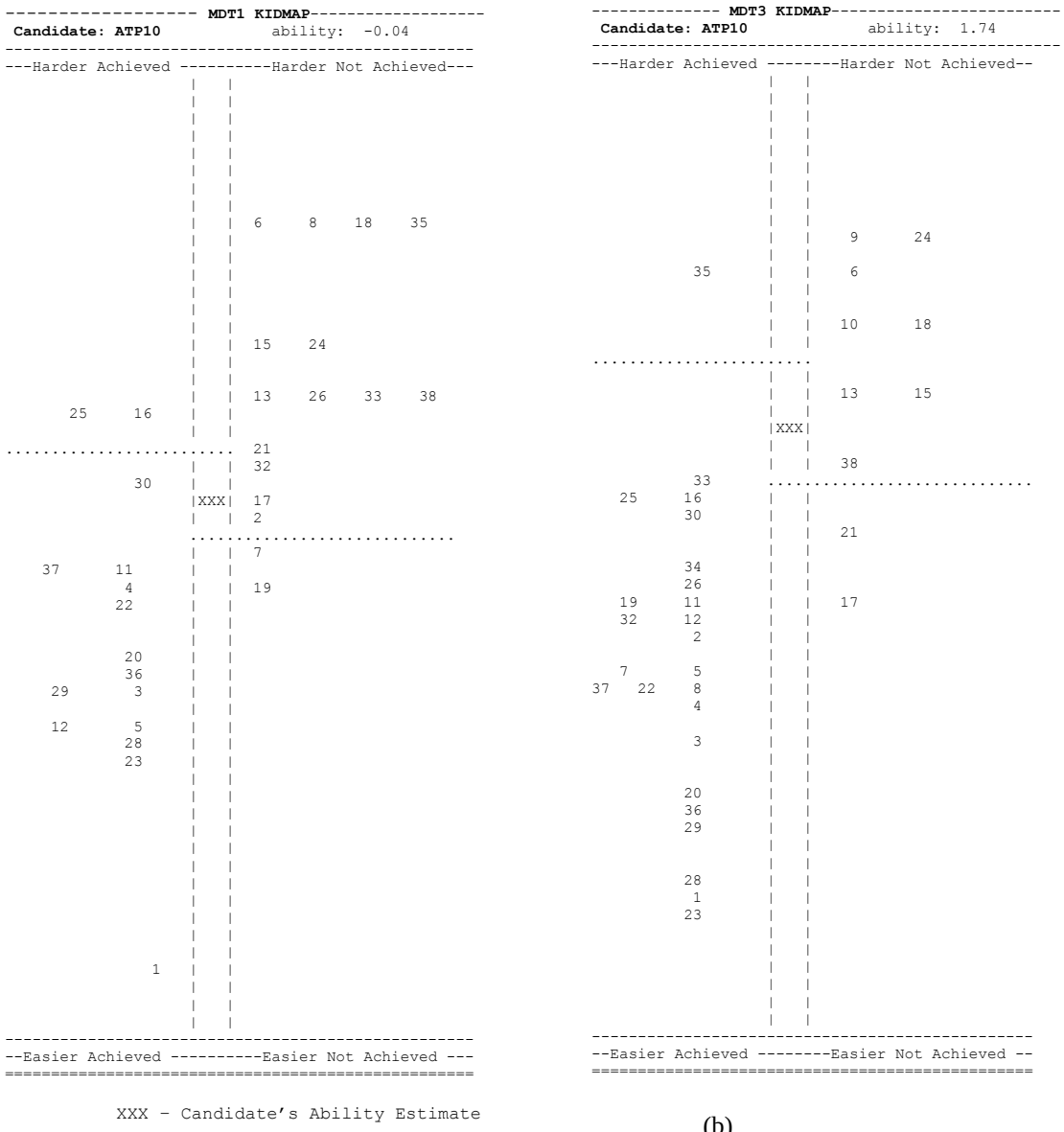
continuum with a XXX (see Figure 3a). Horizontal dotted lines on either side of XXX indicate, to the left the upper boundary (estimate + S.E.) of his ability estimate and to the right the lower boundary (estimate – S.E.) effectively dividing the kidmap into 4 sections. To the left of the vertical dotted lines are all items that ATP10 got correct (16 items) while on the right are all items that were incorrect (16 items). Excluded from the map are the zero-score items for that test.

Graphed in the top-left section are two correct items (Items 25 and 16), which, according to the Rasch model, were calibrated to be relatively harder for ATP10 but which he got correct, indicating special knowledge. Item 25 was on probability of an event while Item 16 was on interpreting a pictograph scale. Within his ability band to the left is another correct item (Item 30), which, according to the Rasch model, means he had average probability of getting correct. In the bottom-right section are other correct items whose estimates were below his ability estimate, and as the model predicted, he was successful with them. These items represented areas of consolidated understanding at the time of the test.

Within his ability band to the right are incorrect items (i.e., Items 21, 32, 17 and 2) that, theoretically, ATP10 had average probability of getting correct, thus representing areas of difficulties for him. Below the lower bound of his ability band are two more incorrect items (Items 7 and 19) that, theoretically, were easier for him. Item 7 was on representing number relationships as a mathematical equation while Item 19 was on evaluation of a rational algebraic expression. These items represented gaps in his algebraic knowledge and skills at the time of the test. In the top-right section of the kidmap are 10 items that were relatively more difficult for ATP10. These items represented knowledge and skills that he would need to develop and improve. The MDT1 zero-score items would also belong in this section as being too difficult for ATP10.

ATP10's MDT3 mathematical performance is captured by his MDT3 kidmap provided in Figure 3b. It demonstrated improvement, since MDT1, in terms of increased ability estimate of 1.74 logits (S.E. 0.48) (MDT1 ability estimate was -0.04 logits [S.E 0.47]). In fact, Cohen's *d* (of 3.80) demonstrated that ATP10's mathematical performance improved by 3.8 standard deviations. That is, the effect of the two teaching blocks on ATP10's mathematical performance was practically large. Areas of improvement (i.e., gains and advancements) were indicated by items that he got incorrect in MDT1 but he was successful with in MDT3. These items included, in decreasing order of difficulty, Item 8 on ordering fractions, Item 35 on ratio of a mixture, Item 26 on area model of an equivalent fraction, Item 33 on club membership using "more than" relationship, Item 32 on fraction of an amount word problem, Item 2 on mentally computing percentage of a two-digit number, Item 7 on describing a number relationship as a mathematical equation, and Item 19 on evaluation of a rational algebraic expression. It appeared that the one-week intervention activities impacted positively on ATP10's understanding, knowledge and skills in working with a number of fraction and algebraic related items.

Figure 3
*Candidate ATP10's MDT1 and MDT3 Kidmaps*




Persistent areas of difficulties were identified by checking whether the MDT1 incorrect items were still present as incorrect items in the MDT3 kidmap. Subsequent results included Item 6 on ordering decimals, Item 18 on total distance traveled by an elevator, Item 15 on area of path within nested geometric shapes, Item 24 on operations with fractions word problem, Item 13 on proportional reasoning word problem, and Item 38 on communication of a pattern.

According to the two kidmaps, the relatively easier items (Items 17 and 21) in the bottom-right section of the MDT3 kidmap which were also incorrect in MDT1, remained incorrect although theoretically, ATP10 had more than average probability of being successful. Item 21 is on solving a linear equation and Item 17 visualisation of similar

triangles. These incorrect responses suggested gaps in ATP10's algebraic and basic geometry knowledge and skills which should be addressed in subsequent teaching blocks.

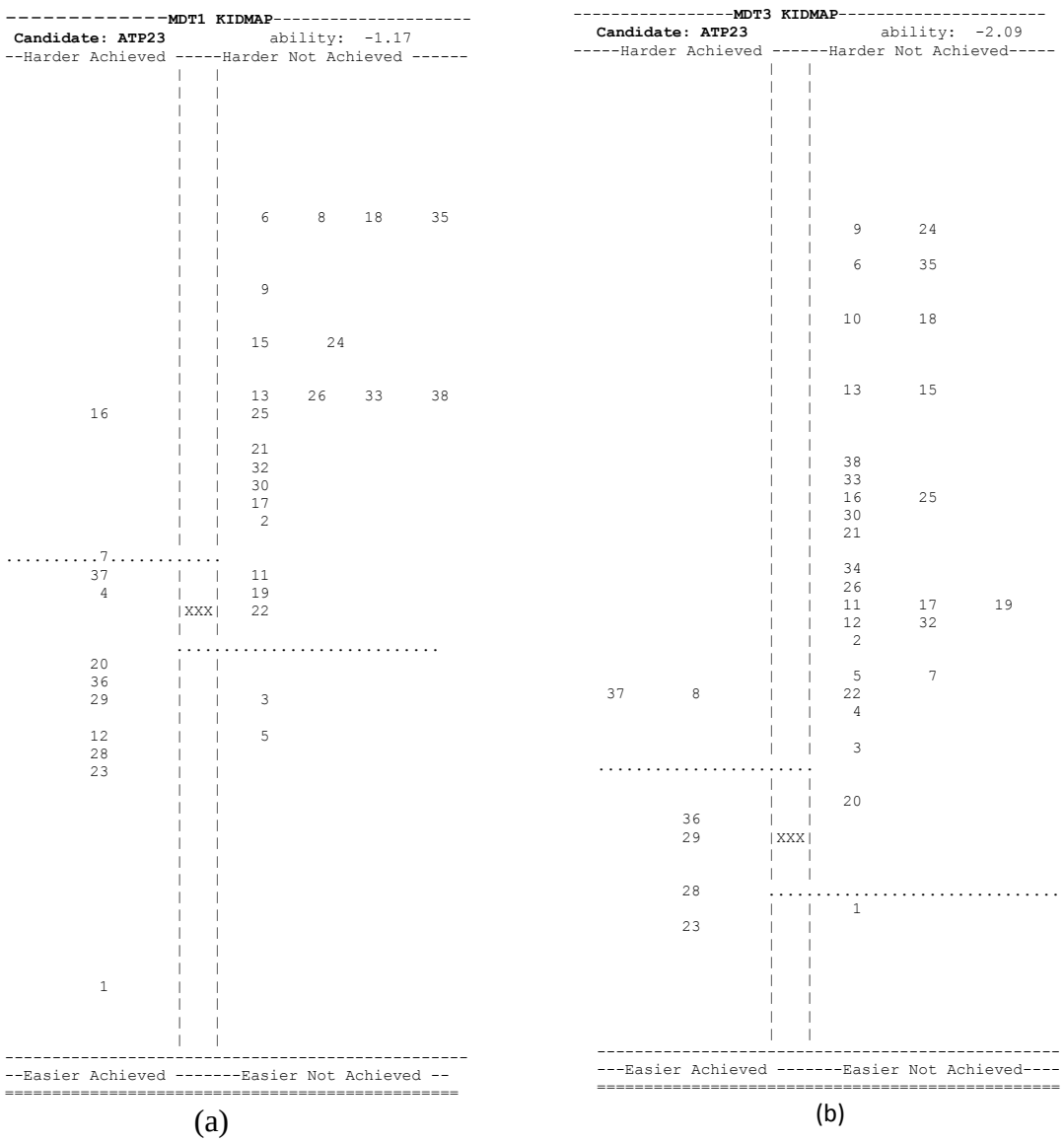
**Mathematical Performance of Candidate ATP23 -** ATP23's overall mathematical performance in MDT1 was calibrated to be -1.17 logits and is located on the logit continuum with a XXX (see Figure 4a). To the left of the vertical dotted lines are all items ATP23 got correct (11 items) while on the right are all items that were incorrect (22 items). Excluded from the map are the zero-score items for that test.

Graphed in the top-left section is one item (Item 16 on interpreting a pictograph scale), which, according to the Rasch model, was calibrated to be relatively harder for ATP23 but which she got correct, indicating special knowledge. Within her ability band to the left are other correct items, in decreasing order of difficulty, (Items 7, 37, and 4), which theoretically, she had average probability of getting correct. Item 7 was on ordering fractions, Item 37 on extending a pattern, and Item 4 on mentally adding unit fractions. In the bottom-left section are 7 other correct items whose estimates were below her ability estimate, and as the model predicted, she was successful with them. These items represented areas of consolidated understanding at the time of the test.

Within her ability band to the right are incorrect items (i.e., Items 11, 19 and 22) that, theoretically, ATP23 had average probability of getting correct, thus representing areas of difficulties for her. Below the lower bound of her ability band are two more incorrect items (Items 3 and 5) that, theoretically, were easier for her. Item 3 was on mentally subtracting 4-digit numbers and Item 5 on mentally multiplying decimals. These items represented gaps in her multi-digit subtraction and decimal multiplication knowledge and skills at the time of the test. In the top-right section of the kidmap are 17 items that were relatively more difficult for ATP23. These items represented knowledge and skills that she would need to develop and improve. The MDT1 zero-score items would also belong in this section as being too difficult for ATP23.

ATP23's MDT3 mathematical performance is captured by her MDT3 kidmap provided in Figure 4b. It illustrated her lack of improvement, since MDT1, in terms of her decreased ability estimate of -2.09 logits (S.E. 0.52) (MDT1 ability estimate was -1.17 logits [S.E 0.48]). In fact, Cohen's *d* (-1.87) demonstrated that ATP23's mathematical performance regressed by about 1.9 standard deviations by the end of the second teaching block. Checking for the status of 22 MDT1 incorrect items in MDT3 showed that 21 items still remained incorrect in MDT3. The only item which demonstrated improvement was Item 8 on ordering fractions which provided evidence of the impact of focused intervention activities during the week. In contrast, correct MDT1 items that became incorrect in MDT3 included Item 16, Item 7, Item 4, Item 20 and Item 12 indicating forgotten knowledge and skills already learnt in the first teaching block.

Figure 4
*Candidate ATP23's MDT1 and MDT3 Kidmaps*




Overall, the display of correct and incorrect items relative to each teacher trainee's existing ability estimate within a test provided diagnostic information to record existing gains and achievements and to inform subsequent intervention activities. The same information over time across tests (e. g., MDT1 and MDT3) provided useful information to identify areas of (a) consolidated understanding and proficiency, (b) gains and advancements, and (c) persistent difficulties. Collectively, the complete series of 3 consecutive kidmaps provided running records upon which to build a more nuanced understanding of an individual student's mathematical performance and profile in terms of strengths and weaknesses by topic area and relevant knowledge and skills.

## Individual Effect Sizes Between MDT3 and MDT1 Mathematical Performances

To illustrate the gains and advancement achieved by each TT as measured by the first and third diagnostic tests, the practical difference between individual performances was determined for all TTs in terms of Cohen's *d*. The results are provided in Figure 5.

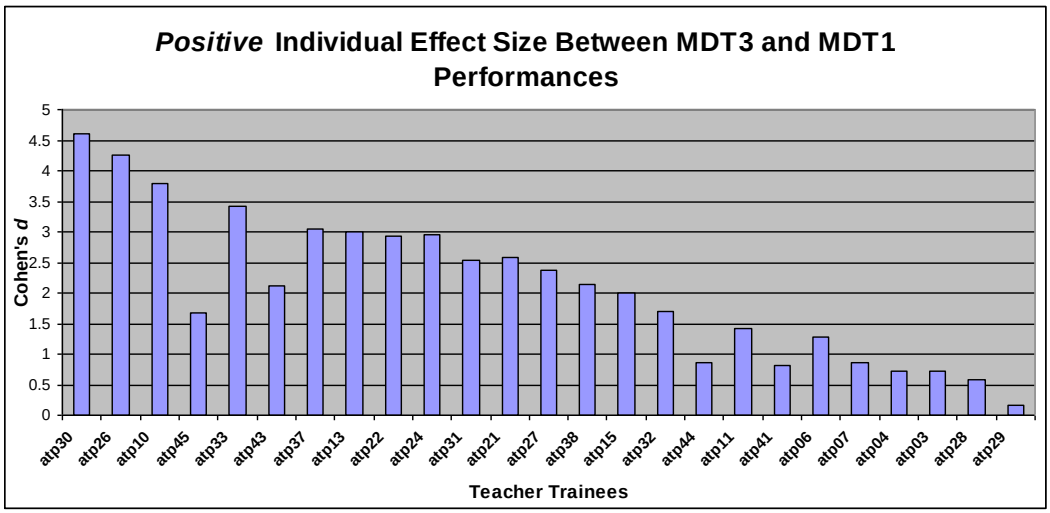
Of the 35 TTs that had MDT1 and MDT3 ability estimates, 71% (25/35) had positive effect sizes (see Figure 5a) including ATP10. The effect sizes ranged from 0.16 to 4.6 implying that the effect of the intensive teaching blocks on these TTs' mathematical performances ranged from 0.16 to 4.6 standard deviations. In contrast, 29% (10/35) had negative effect sizes (see Figure 5b) including ATP23. The negative effect sizes ranged from -2.08 to –0.04 indicating that these TTs did not advance in their learning, instead they regressed. For both categories of TTs, more details about their areas of gains and advancements, consolidated understanding and proficiency and gaps in their knowledge and skills may be extracted from their kidmaps as was done for the two TTs ATP10 and ATP23 in this paper. During the ADEP program itself, individual copies of 3 kidmaps with explanatory notes were provided during class to all TTs at the beginning (MDT1 kidmap) and end (MDT2 kidmap) of the second teaching block and in a special wrap-up session one week later (MDT3 kidmap).

## Discussion

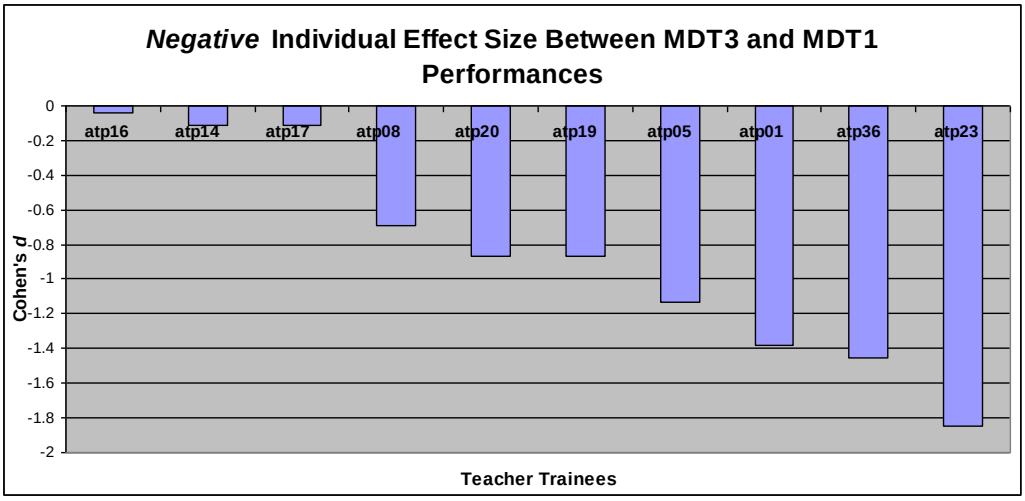
The study monitored the changes in TTs' mathematical performance and examined the nature of the mathematics knowledge and skills that they employed to solve items in the diagnostic tests. The Rasch Measurement Model underpinned the analysis of TTs' responses (after coding them correct, incorrect or blank) to calibrate ability and difficulty estimates. In addition, information extracted from the Quest generated outputs (item analysis data, variable maps and kidmaps) facilitated the identification of areas of advancement and gains and those that were still problematic.

On the bases of variable maps, item analysis data and kidmaps, detailed analyses of TTs' mathematical performance helped researcher/teachers more effectively target topic areas (content, knowledge and skills) to improve in subsequent instruction. For example, receiving timely feedback through individual kidmaps helped both teachers and individual students identified areas that needed further work and improvement. Quest-generated item analysis information also provided useful information on which items tended to have the largest disparities between success and error percentages and identifying topics most in need of improved teaching and further development. Large disparities between success and error rates suggested that TTs have had relatively few opportunities to solve more challenging (multi-step) mathematics problems and to learn about fractions, percentage and measurement in meaningful ways as evidenced by the zero-score items and the "most difficult" items located in the top right section of variable maps and kidmaps. Overall, TTs' ability measures throughout the first three diagnostic tests charted individual's performances from which the teachers could determine whether or not an individual was progressing as expected.

Figure 5
*Effect Sizes between Individual Performances in MDT1 and MDT3*




(a)




(b)

A drive for excellence to achieve success with all or most items in the right column of their individual kidmaps to obtain high ability estimates indicating high performance, led TTs to seek large gains or advancement in their subsequent test performances as demonstrated by APT10 and others with positive effect sizes between the first and third diagnostic tests.. Discrepancies in mathematical performances mirrored discrepancies in opportunities to be exposed to solving items such as those in the diagnostic tests (e.g., challenging multi-step word problems) that TTs from different mathematical backgrounds experienced prior to participation in the ADEP program. For example, where test items were presented as word problems, the TTs' language comprehension as well as their mathematical skills were involved in the responses that were generated. Most likely, difficulties with language comprehension prevented TTs from accessing the mathematics embedded in the descriptions; for example, Item 14 and Item 27 which were consistently zero-scored across the 3 tests. The context within which this ADEP program was

undertaken (i.e., compressed period of time) may not have suited all TTs. However, conducting this research and doing Rasch analyses of test results can shape the climate of opinion about the accelerated program and inform national and university policies about the feasibility of such programs for the training of future Samoan primary teachers. Evidence-based information about the gains and advancements (or lack of it) of each teacher trainee or which item types and mathematical topics were most in need of targeting was important for the researcher/teachers to draw from as they designed focused interventions in subsequent teaching blocks (see Lubienski, 2008). The continuing support provided to the TTs through focused intervention activities and timely feedback through kidmaps may have contributed substantively to the impact demonstrated by individual effect sizes between MDT1 and MDT3 mathematics performances.

The work that is presented in this paper focused on the repeated measurement of TTs' mathematical ability within the context of their accelerated teacher education program to provide evidence of their on-going mathematical competence. Althoug, teachers' mathematical content knowledge in itself is not strongly linked to student achievement (Ball, Lubienski, & Mewborn, 2001; Wilson, Floden, & Ferrini-Mundy, 2002), it is clearly the case that mathematical content knowledge is essential for teachers to lead mathematical discussions and is subsequently conceptualized to form the basis of other types of knowledge that play a substantial role in teachers' practices and the learning opportunities they design for students (e.g., Fennema et al., 1996; Fennema, Franke, Carpenter, & Carey, 1993; Hill et al., 2005; Hill et al., 2004). Of particular note are the overlaps between pedagogical content knowledge (Shulman, 1986) and specialized content knowledge (Ball, Thames, & Phelps, 2008; Hill et al. 2008) as defined and conceptualized in the literature to include content knowledge of the subject matter. The empirical findings from this study contribute to the literature on teacher education in that, with focused intervention on particular areas of difficulties, teacher trainees' mathematical understanding and competence can improve over time.

## Conclusion

The conclusions of this study are based on the data presented and subsequent findings are formulated as answers to the paper's two focus questions: (1) *What are TTs' base line mathematical abilities and identified areas of difficulties early in their ADEP program?* (2) *What are TTs' mathematical abilities, areas of improvement and areas of on-going difficulties by the end of the first two teaching blocks?*

For the first focus question, the TTs' baseline mathematical abilities, as calibrated using the Rasch Measurement Model with the Quest software, were as displayed on the MDT1 variable map for the group and on individual MDT1 kidmaps for each teacher trainee. It was found that the TTs' baseline estimates were mainly low (average ability was below zero) with more detailed information about individual problematic areas provided by kidmaps. The latter provided individual feedback in terms of overall performance, strengths and areas requiring further development. A synthesis of information extracted from variable maps, item analysis data and kidmaps enabled the identification of problematic areas requiring further development. These areas included multi-step word problems, ordering fractions and decimals, probability, and basic algebra, geometry and measurement knowledge and skills.

For the second focus question, by comparing mathematical performances across the 3 tests, it was found that the 12-month break led to some loss of first learning with group average ability remaining more or less the same. However, with focused intervention over one week, the third diagnostic test's results demonstrated a moderate effect size when compared with baseline estimates. In terms of individual performances, the majority of TTs demonstrated positive effects ranging from small to very large effect sizes. Whilst there was still room for further improvement to achieve proficiency level beyond 2.0 logits, the data presented and analysed provided empirical evidence to support the gains and advancements made by these TTs at this point of the ADEP. Areas of improvement were noted with ordering fractions, probability and some word problems but multi-step word problems and ordering decimals continued to be persistent areas of difficulties.

## Implications

Findings grounded on the actual responses of TTs to the diagnostic test items over time provide valuable information to (a) the researcher/teachers to further inform the adaptation and modification of learning activities in subsequent teaching blocks, and (b) individual teacher trainees to track and monitor their progress over time during the duration of their ADEP program. Tracing the development of each TT's progress through their ability estimates and consecutive kidmaps over time provides empirical evidence of their own gains and advancements or lack of it. The linking of their mathematical abilities and kinds of items they got correct and incorrect over time provides personally useful diagnostic information to further guide subsequent focused intervention activities for the rest of their ADEP program and later on for their own self-learning between teaching blocks. The conclusions imply that critical skills involved in solving challenging word problems, ordering and operations with fractions, identifying relevant information from complex area diagrams and solving probability related items are areas that require explicit development and improvement especially for future primary teachers.

In addition to contributing to the scant research on teacher education in Samoa, we gain the added bonus of studying teacher trainees whose mathematical content knowledge was initially limited thus allowing a clearer picture of the kinds of progress that teacher trainees achieve within a compressed period of time. Doing so provided evidence-based findings to inform further development of both the accelerated and normal 2-year Diploma in Education (Primary) Programs so that graduates exit better equipped to cope with the challenges of the new primary mathematics curriculum.

## References

Adams, R. J., & Khoo, S. T. (1996). *QUEST*: *Interactive item analysis.* Melbourne: Australian Council for Education and Research.

Afamasaga-Fuata'i, K., Meyer, P., & Fili-Falo, N. (2010). Mathematically speaking: Where are we now? Where are we going? Snapshots of some secondary and post-secondary students' mathematics performance. In L. I. Fuata'i & T. Lafotanaoa (Eds.) *Measina a Samoa* p*roceedings of the measina a Samoa IV conference 15-17 December 2008* (Volume 4, 125-152). National University of Samoa.

Afamasaga-Fuata'i, K., Meyer, P., & Falo, N. (2008). Assessing primary preservice teachers' mathematical competence. In M. Goos, R. Brown, & K. Makar (Eds.), *Navigating*

*currents and charting directions. Proceedings of the 31st annual conference of the mathematics education research group of Australasia* (Volume 1, pp. 43-49). University of Queensland, Australia: MERGA.

Afamasaga-Fuata'i, K., Meyer, P., & Falo, N. (2007). Primary students' diagnosed mathematical competence in semester one of their Studies. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice. Proceedings of the 30th annual conference of the mathematics education research group of Australasia* (Volume 1, pp. 83-92). University of Tasmania, Australia: MERGA.

Afamasaga-Fuata'i, K., Meyer, P., Falo, N., & Sufia, P., (2007). Future teachers' developing numeracy and mathematical competence as assessed by two diagnostic tests [AARE's website]. Retrieved from http://www.aare.edu.au/06pap/afa06011.pdf.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives, 9*(1), 95-104.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model. Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*(2), 125-144.

Australian Association of Mathematics Teachers (AAMT). (2007). *AAMT standards for excellence in teaching mathematics in Australian schools* [online site]. Retrieved from http://www.aamt.edu.au/standards, on October 6, 2007.

Ball, D. L. (1993). With an eye on the mathematical horizon: Dilemmas of teaching elementary school mathematics. *Elementary School Journal, 94*, 373–397.

Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *Elementary School Journal, 90,* 449-466.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83–104). Westport, CT: Ablex.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, DC: American Educational Research Association.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*, 389–407.

Borko, H., & Putnam, R. T. (1996). Learning to teach. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 673–708). New York: Simon & Schuster Macmillan.

Briars, D. (2001, March). *Mathematics performance in the Pittsburgh public schools*. Paper presented at a meeting of the Mathematics Assessment Resource Service, San Diego, CA.

Briars, D. J., & Resnick, L. B. (2000). Standards, assessments—and what else? The essential elements of standards-based school improvement (CSE Technical Report 528) [online site]. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH528.pdf

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*, 311–329.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second ed.). NY: Lawrence Erlbaum Associates.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*, 403–434.

Fennema, E., Franke, M. L., Carpenter, T. P., & Carey, D. A. (1993). Using children's mathematical knowledge in instruction. *American Educational Research Journal, 30*, 555–583.

Heaton, R. M. (2000). *Teaching mathematics to the new standards: Relearning the dance*. New York: Teachers College Press.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*, 372–400.

Hill, H., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105*, 11–30.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

Lubienski, S.T. (2008). On "gap gazing" in mathematics education: The need for gaps analyses. Research Commentary. *Journal of Research in Mathematics, 39*(4), 35-356.

Nathan, M. J., & Knuth, E. J. (2003). A study of whole classroom mathematical discourse and teacher change. *Cognition and Instruction, 21*, 175–207.

National Council of Teachers of Mathematics (NCTM). (2007). *Executive Summary. Principles and Standards for School Mathematics* [on-line site]. Retrieved from http://standards.nctm.org/document/chapter3/index.htm.

New South Wales Board of Studies (NSWBOS) (2002). *K-6 Mathematics Syllabus*. Sydney, Australia: NSWBOS.

Piaget, J. (1972). *The psychology of the child*. New York: Basic Books.

Punch, K. (2009). *Introduction to Research Methods in Education.* Sage Publications Ltd, London.

Rasch Analysis (2005). *What is Rasch Analysis?* Retrieved online from http://www.rasch-analysis.com.

Rasch, G. (1980). Probabilistic Models For Some Intelligence And Attainment Test. (expanded version). Chicago: The University of Chicago Press.

Samoan Ministry of Education, Sports & Culture (SMESC). (2010). ESP 2 Project Documentation. SMESC.

Samoa's Ministry of Education, Sports & Culture (MESC). (2008). Educational Statistical Digest 2008, Part 2, pp. 3-4 [online site] Retrieved from http://www.mesc.gov.ws/pdf/edu_stats_digest_2008.pdf.

Sherin, M. G. (2002). A balancing act: Developing a discourse community in a mathematics classroom. *Journal of Mathematics Teacher Education, 5*, 205–233.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10*, 313–340.

Thalheimer, W., & Cook, S. (2002, August). *How to calculate effect sizes from published research articles: A simplified methodology* [on-line site] Retrieved from http://work-learning.com/effect_sizes.htm.

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

Vygotsky, L. (1978). *Mind and Society.* Cambridge, MA: Harvard University Press.

Williams, S. R., & Baxter, J. A. (1996). Dilemmas of discourse-oriented teaching in one middle school mathematics classroom. *The Elementary School Journal, 97*, 21–38.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2002). Teacher preparation research: An insider's view from the outside. *Journal of Teacher Education, 53*, 190–204.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

## About the Author

**Karoline Afamasaga-Fuata'i** (k.fuatai@nus.edu.mail) is the Founding Professor of Mathematics and Mathematics Education at the National University of Samoa, Samoa. Her current research interests include assessment of students' mathematics abilities and attitudes using diagnostic testing and attitudinal questionnaires respectively, and the Rasch Model for data analysis. Karoline also analyses qualitatively students' open responses from their authentic mathematical investigations and innovative use of meta-cognitive tools such as concept mapping and vee diagrams.

# Appendix 1

| Item | Brief Item descriptions |
|---|---|
| 1 | Mental computation – multiplication of 1/2-digit numbers |
| 2 | Mental computation – percentage of a whole number |
| 3 | Mental computation – 4-digit subtraction |
| 4 | Mental computation – addition of unit fractions |
| 5 | Mental computation – multiplication of 1-decimal place numbers |
| 6 | Ordering decimal fractions |
| 7 | Number relationship described as a mathematical expression |
| 8 | Ordering fractions |
| 9 | Average speed word problem |
| 10 | Probability word problem |
| 11 | Fraction of a number problem |
| 12 | Reading a bar graph |
| 13 | Proportional reasoning word problem |
| 14 | Describing a relationship between two quantities as a mathematical expression |
| 15 | Nested areas – shape within another shape |
| 16 | Pictograph scale |
| 17 | Similar triangles visualisation |
| 18 | Elevator word problem |
| 19 | Evaluation of an algebraic expression |
| 20 | Missing angle in a quadrilateral |
| 21 | Solving a linear equation |
| 22 | Equivalent fractions |
| 23 | Estimate of an angle measure |
| 24 | Operations with fractions – word problem |
| 25 | Probability of an event |
| 26 | Area model of a fraction |
| 27 | Area of rectangle & ratios |
| 28 | Index notation |
| 29 | Similar triangles – proportional reasoning |
| 30 | Average speed word problem |
| 31 | Average weight of crystal |
| 32 | Fraction of a number and balance left |
| 33 | Club membership using "more than" relationship |
| 34 | Shaded area of a geometric shape |
| 35 | Ratio of a mixture |
| 36 | Pattern extension from diagrams |
| 37 | Pattern extension from table of values |
| 38 | Communication of pattern from table |

Educational Measurement and Evaluation Review (2011), Vol. 2, 62-76


http://pemea.club.officelive.com/ EMEReview.aspx

# Determining Experts and Novices in College Algebra: A Psychometric Test Development and Analysis Using the Rasch Model (1PL-IRT)

**Jonathan V. Macayan**
*SLHS-Department of Psychology*
*Mapua Institute of Technology*
*Manila, Philippines*

**Bernardino C. Ofalia**
*College of Human Development*
*Pamantasan ng Lungsod ng Maynila*
*Manila, Philippines*

**Abstract** This study involved the development and examination of the psychometric properties of a proposed measure (College Algebra Diagnostic Test) which intends to test the ability of students in College Algebra. The 15-item instrument was constructed, subjected to content validation, and subsequently tested with college students (*N*=180) in a private HEI in Manila, Philippines. Test results were psychometrically analyzed using the Rasch model (1PL-IRT). The 1PL-IRT statistical outputs were subsequently analyzed particularly on reliability and item difficulty indices. It was found that the IRT results on item reliability was very high (.97, RMSE = .22) while the person reliability (.55, RMSE = .72) was constrained on a moderate level due to a high degree of standard error. In the analysis of item difficulty, IRT categorized 9 out of 15 items as *easy* and 6 as *difficult*. The data also supported the assumption of unidimensionality (1.09) of the construct being measured by the items. Goodness of fit was also provided (INFIT =1.00; OUTFIT = .92). Finally, the study recommended further investigation on the psychometric properties of the test using advanced IRT (2PL or 3PL) models.

**Keywords:** *Psychometric properties, College Algebra, Rasch Model*

## Introduction

The differential performance of experts and novices has long been a focal area of investigation in research in education. These researches are basically oriented towards the conceptualization of appropriate learning strategies that would make learners of different levels of ability benefit from formal education in various learning domains.

The term *diagnosis* is a common expression in clinical parlance. It usually refers to the initial step of clinicians in determining the underlying disorder or pathology as manifested through symptoms. This helps the practitioner in deciding on the appropriate treatment plan or rehabilitation program for a specific condition of a patient. Diagnosis

may assume another definition beyond the clinical perspective; in the field of education for instance, diagnosis implies an instructional description in which assessment results provide information about students' mastery of relevant prior knowledge and skills within the domain as well as the perceptions about the course.

Diagnosis is a vital aspect of instructional decision-making. It may serve as an instrument to determine students who may be at-risk for failure and it helps educators in the delivery of carefully designed supplemental interventions. Further, diagnosis provides valuable information about students' persistent misconceptions in the targeted domain (Ketterlin-Geller & Yovanoff, 2009).

Traditionally, educators principally focus on summative assessment which reflects whether or not the students learned what they are supposed to learn; this is usually gauged through achievement tests. Unfortunately however, testing at the end of the course (like in the case of summative tests) does not provide facilitative measures to help improve the performance of learners in a specific subject domain since it only assesses the accumulated knowledge and not the prior knowledge of learners which is considerably an important predictor of their achievements and success in learning. Diagnostic tests (unlike achievement tests) focus on the facilitative mechanism of learning; thus, these kinds of tests are categorized under *formative assessment*. Diagnostic testing can help educators achieve the important goal of education which is to improve students' learning.

## Diagnosis of Mathematical Ability

Mathematics has been considered by many as one of the toughest subject matter in school. Studies involving math identified several factors that contribute to the complexity of the subject matter such as the fear factor or the negative anticipation of difficulty towards the subject on the part of the learner, and the poor prior training received from earlier schooling. Adding to the inherent complexity of mathematics are the inappropriate instructional approaches or techniques used by some teachers in facilitating lessons in math classes.

One of the most crucial steps to ascertain effective learning in mathematics that warrants serious consideration from teachers is the assessment and diagnosis of students' prior knowledge or level of expertise in the domain. By having knowledge about the students' levels of expertise, teachers can devise more appropriate techniques that will effectively facilitate learning among students of varying levels of ability. Several studies emphasized that different levels of expertise would require different strategies in order to ensure learning (Kalyuga, 2007; Rittle-Johnson & Kmicikewycz, 2008). Thus, a particular teaching strategy may be found effective for novice learners but may not be beneficial to learners who have some experience in the domain and vice-versa. Based on Rittle-Johnson, Star, and Durkin (2009), novice learners experience difficulty in coping with the cognitive demands of learning mathematics since math lessons can easily overload their working memory, as they must deal with enormous and new elements of information at once. On the contrary, a more experienced learner can easily adapt to the cognitive requirements in learning mathematics since they can use their existing knowledge structures to interpret and integrate new information without overloading their working memory.

An example of an attempt to diagnose mathematical ability in the tertiary level is the study by Ramos (2008) in which a diagnostic test was constructed and evaluated. The primary purpose of this study was to develop and design an instrument that can be used to assess college students' preparedness in a College Algebra course. 73 items were used and pilot tested with 595 students. The results of reliability testing showed that the instrument is reliable at a 0.05 significance level. Concurrent and predictive validity testing likewise showed that the instrument is valid. Based on the decided cut-off score, students who scored at least 56 (out of 73) in the test were found to be more likely to pass a College Algebra course than those who scored less than 56.

## Algebra: Experts vs. Novice

In the Philippine education setting, Algebra is usually the first mathematics course that students encounter in the tertiary level; this course usually builds up the foundation of college level mathematics. In mathematics-oriented college degree programs (i.e., engineering), Algebra is considered a crucial subject-domain; thus, it is usually assigned as the pre-requisite to higher math courses. Based on observation, many students are struggling with the course resulting to repeated failures and consequently result to setbacks in progressing through the curriculum. This observation may have something to do with the difficulties that students encounter in the transition from arithmetic to algebraic thinking (Knuth, Stephens, McNeil, & Alibali, 2006). For example, in arithmetic symbolic language, the equals sign (=) is usually conceived as an indicator of the result or answer to the problem. A student who would tend to fixate on this notion of equals sign would most likely commit serious error in the problem solution in Algebra since equals sign in algebraic notation may also infer equivalence of two phrases (2x = x2...) and not only an indicator of answer to the problem (Baroody & Ginsburg, 1993). This error is conceptual in nature since the inability of a student to provide correct answer to the problem is due to his inexact understanding of the functions of algebraic symbols and terms. In addition to conceptual inaccuracy, another possible reason why students struggle in Algebra is the lack of knowledge in the procedural aspect. According to Lerch (2004), students often use incorrect procedures when learning Algebra and this inhibits accurate solutions to the problems; such difficulty in the procedural aspect has been attributed to the inadequacy of student's knowledge of the problem features.

These problems in conceptual and procedural knowledge in Algebra would distinctively separate experts and novices. In usual cases, novice learners in the subject would have trouble coming up with the right problem solutions since they lack the necessary understanding of the fundamental concepts and appropriate strategies involved in solving algebraic problems. In a more cognitive sense, this phenomenon has been associated with the memory transformations involving two specific memory functionalities such as the working and long-term memory (Sebrechts, Enright, Bennet, & Martin, 1996). The capacity of working memory can vary with the level of expertise of individuals; expert learners who have a good deal of schema stored in the long-term memory may have little difficulty solving a quadratic equation compared with novice learners because the former may process the complex equation as a single chunk while for the latter it may be processed as multiple separate chunks.

As cited in Magno and Ouano (2010, p.136), to become a good problem solver in mathematics, one must develop a base of mathematics knowledge. Thus, poor foundation knowledge in mathematics may translate into difficulty in giving correct problem solutions. The same holds true in performing solutions to algebraic problems; novices who are not familiar or may not have accurate understanding of algebraic notations, terms, and symbols may have difficulty deriving a correct answer to the problem. Experts on the other hand may struggle less since they are equipped with the essential knowledge or conceptual understanding of these algebraic elements. According to Schoenfeld and Hermann (1982), a fundamental difference between novices and experts in their approaches to problem solving in math is that a novice learner attend to surface features of problem while an expert learner categorized problems on the basis of the fundamental principles involved.

Schoenfeld (1985) further explained the fundamental basis of mathematical thinking (particularly that of problem solving tasks) by conceptualizing a framework for the analysis of mathematical behavior. In his framework, he specified four categories of knowledge and behavior which caused learners to succeed or fail in their attempts to solve mathematical problems: (1.) *resources* – are the body of knowledge that an individual is capable of bringing to bear in a particular mathematical situation; (2.) *heuristics* – are rules of thums for effective problem solving; (3.) *control* – deals with the question of resource management and allocation during problem solving attempts; and (4.) *belief system* – one's mathematical world view or the perspective with which one approaches mathematics and mathematical tasks.

Among the four categories explained by Schoenfeld (1985), the *resources* significantly distinguishes expert and novice learners in mathematical problem solving. A more experienced learner may likely to succeed even in a highly complex problem situation since they have sufficient cognitive resources (i.e. schema, experience, etc.) to use in solving problems, unlike novices who do not yet have an organized sense of knowledge and experience which can be used to facilitate during the problem solving tasks. In this contention, Schoenfeld specified that the resources is not a standalone feature which elaborates the learner's mathematical behavior; each category contributes with another. Thus, an expert learner is not only characterized by having sufficient experiential and cognitive resources but by his performance in problem solving situations is also directed and typified by his strategies (heuristics), belief system, and management of resources (control).

The facts about the difference between experts and novices in their approach to learning necessitate serious actions from educators. But this cannot be done without assessing of students' learning potentials through careful diagnostics. Thus, the present study intends to design a diagnostic instrument that can be useful in screening the learner's level of expertise in mathematics particularly in College Algebra. This study hopes to contribute to education literature by psychometrically designing, analyzing and validating a diagnostic test which can help college instructors in mathematics to distinguish varying levels of college students' mathematical ability which in turn can be utilized in employing appropriate teaching strategies that can be beneficial to learners of diverse abilities.

## The Psychometric Analysis: Validity and Reliability

The most important intention in this study is to come up with a valid and reliable instrument which can distinguish high ability (expert) and low ability (novice) problem solvers in College Algebra. To realize this goal, the Rasch model (also known as 1 Parameter Logistic Model - 1PL-IRT) was applied. The following section explains the psychometric approach used and some test development studies conducted using the model.

Psychometric test analyses are used in determining two of the most important test's indices: Item difficulty and item discrimination. Item difficulty distinguishes whether an item in a test is easy or difficult. Generally, a test item is easy if a large number (%) of test takers are able to answer it correctly, and conversely it is difficult if only a few (%) are able to answer it correctly. Item discrimination distinguishes whether an item in a test is good or poor in reference to estimating the difference between test-takers of high and low ability in getting the correct answer to an item. Usually, test-takers who are of high ability should be able to answer test items correctly (especially difficult ones) while those who have low ability would generally not be able to answer items correctly. High and low ability is principally based on the overall performance of examinees on a test and is primarily based on their total score (Payne, 1992).

The Classical Test Theory has been the leading model for analyzing and developing several kinds of tests; in fact, it has dominated the area of test analysis until at least before the advent of IRT in 1970's. It is regarded as the *True Score Theory* since the model is based on the assumption that a test-taker has an observed score and a true score (Magno & Ouano, 2010). The observed score of a test-taker is usually seen as an estimate of the true scores of that test-taker and (plus/minus) some unobservable measurement error (Hambleton & Swaminathan, 1985). An advantage of CTT is that it relies on weak assumptions and is relatively easy to interpret. However, CTT has been heavily criticized in its basic assumption of true score, since the true score is not an absolute characteristic of a test-taker because it depends on the content of the test. Another criticism is that the items' difficulty may vary depending on the sample of test-takers who take a specific test. Therefore, it is difficult to compare test-takers' results between different tests. Due to these limitations, the modern approach called the Item Response Theory (IRT) came into place.

IRT (also known as latent trait theory or strong true score theory) is a fairly new framework of test analysis; it makes stronger assumptions as compared to CTT. The modern framework is based on item analysis as it considers the chances of test-takers (with given ability) in getting particular test items correctly or incorrectly (Kaplan & Saccuzzo, 1997). The Rasch model which was developed by Rasch (1960) having its main motivation of eliminating references to populations of examinees in analyses of tests is one of the variants (2PL and 3PL) of IRT approaches. As an IRT based framework, the Rasch model primarily uses the difficulty parameter as the single and relevant dimension to analyze test characteristics; thus, the model is categorized as one parameter logistic (1PL-IRT).

Since the beginning of the 1970's, IRT has more or less overthrown CTT and is now the most trusted framework being used in the field of test analysis (Hambleton & Rogers, 1990). One of the most important conjectures in IRT is that the true (latent) ability of a test-taker is not dependent on the content of a test. For IRT, it is also essential to assume unidimensionality; that is, the items in a test should measure a single latent ability.

Based on this framework, test-takers with high ability should have a high probability of answering an item correctly while test-takers with low ability will not. In addition, IRT assumes that it does not matter which items are used in order to estimate the test-takers' ability. This assumption allows test evaluators to compare test-takers' result even if they take different forms or versions of a test (Hambleton & Swaminathan, 1985).

Several studies have been conducted to compare the effectiveness of CTT and IRT in test analysis. Most of these studies have proven the advantage of IRT analysis over CTT while some studies have proven that the two frameworks do not differ significantly in determining psychometric indices of tests.

In the study of Magno (2009), he compared and demonstrated the difference between CTT and IRT based on estimates of item difficulty, internal consistency values, variation of ability, and measurement errors using a chemistry test for junior high school students. He found that IRT (One Parameter) estimates of item difficulty are stable and do not change across two samples who took two forms of Chemistry test, unlike in CTT analysis which resulted in inconsistencies. In the analysis of test reliability, IRT provided very stable internal consistencies across samples, something which CTT failed to demonstrate. Further, IRT had significantly less measurement errors than the CTT approach. These findings strongly indicate the difference between the two approaches pertaining to sampling and test analysis; that IRT displayed more effective and accurate analytic outputs than CTT.

The study of Wiberg (2004) aimed at examining which IRT model is the most suitable for use when evaluating the theory test in the Swedish licensure test. A sample of 5,404 test-takers who took one of the test versions of the Swedish theory driving license test was used to evaluate the test results. The study concluded that the estimates are valid for both CTT and IRT. The findings suggested that IRT gives valuable information about a test-taker's true knowledge. IRT has the advantage that the estimates of the item parameters are independent from the sample that has been used. This advantage is especially useful when reusing a test a number of times. From the ICC it is clear how the items work, and which ability a test-taker has that performed well on each item. The TIF and the standard error give a measure of the amount of information that is obtained from the test about a test-taker depending on the test-takers ability level. Finally, if both CTT and IRT are used when evaluating items, different dimensions of information are obtained since both CTT and IRT add valuable information about the test.

The present study involved a thorough psychometric test analysis using the Rasch model (1PL-IRT). Through the IRT, the study's aim of testing the proposed measure's validity, reliability, difficulty, and dimensionality was realized. The model was used to establish a valid and reliable measure that can estimate the differential ability of test-takers with an intention of diagnosing the level of expertise (Expert vs. Novice) of students prior to taking a College Algebra course. Subsequently, the results of diagnosis using the proposed instrument can aid teachers in instructional decision-making.

# Method

## Participants

The study involved college students (*N*=180) who completed the constructed College Algebra Diagnostic Test. The participants are composed of freshmen and sophomore engineering students of a private HEI in Manila; they were conveniently selected from the classes of various engineering sections who are currently taking General Psychology course. The freshmen and sophomore classes were necessary in this study to establish a wider range of possibilities of including experts and novices on the domain that the diagnostic test intends to measure.

## Instruments

The test (Diagnostic Test for College Algebra) intends to measure the levels of students' prior knowledge in College Algebra domain. The test was initially composed of 22 items with multiple choice (A – D) response formats. Prior to content validation, an initial consultation was held with 3 expert validators where the purposes and coverage of the test were discussed. The first draft of the instrument was subjected to thorough item review. After the content (item) validation, 15 items were recommended for inclusion in the test while 7 items were rejected due to either inappropriateness to the level of prospected test-takers or the items' relative complexity due to overlapping concepts with other fields of mathematics (i. e., trigonometry). Some items in the test require a single answer-solution while some necessitate multiple answer-solutions. Items with single answer-solution only require one algebraic operation; these items pose only a single problem to be answered. On the other hand, items with multiple answer-solutions may require two or more algebraic operations to answer the item (see Table 1.0 for sample items).

After the necessary revisions were incorporated in the test, it was pre-tested to 4 college students (2= engineering majors, 2=social science majors) from a private engineering institution in order to determine the length of time needed by test-takers in answering the entire instrument. The average time was calculated (M= 59.30); thus, the allotted time of 60 minutes was included in the test instruction.

## Procedures

The finalized instrument was administered to college students (*N*=180) in a private HEI in Manila. The participants were composed of freshmen and sophomore engineering students who are currently taking General Psychology during the term of test-administration. A letter was submitted to 3 faculty members who are in-charge of 5 classes of engineering students requesting approval for the sessions of test-administration and debriefing. The instrument was personally administered by the proponent so as not to create variations on administration which may have negative implications on test-takers performance. For each class, the entire test administration (including reading of instructions and entertaining queries) covered 70 minutes. After the completion of the test, the participants together with their instructors were debriefed about the purpose of the study.

Table 1
*Sample Items in College Algebra Diagnostic Test*

| No. of Answer/Solution | Sample Item |
|---|---|
| Single Answer/Solution | Item No.3. *At a certain time, Edward notices that his watch reads* ***a*** *minutes after two o'clock. Fifteen minutes later, it reads* ***b*** *minutes after three o'clock. He is amused to note that* ***a*** *is six times the value of* ***b****. What time was it when he looked at his watch for the second time?*<br><br>a. 3:09 b. 3:08 c. 3:10 d. 3:11 |
| Multiple Answers/Solutions (2 operations) | Item No.11. *Colette bought 5 kilos of mangoes at P50.00 a kilo and 2 kilos of grapes at P180.00 a kilo. How much change should she get if she gave a P1,000.00 bill? What is the correct equation that will represent the amount she has to pay if she increases her order for mangoes and grapes?*<br><br>a. 540, 180x + 50y b. 450, 40x + 18y c. 460, 180x + 40y d. 540, 50x + 180y |
| Multiple Answers/Solutions (3 operations) | Item No. 9. *After walking a distance of 3 kilometers, Joshua ate and rested for a while and continued walking at the rate of 5kms per hour. How far has he gone after 4 hours? After 6 hours? What is the correct equation that will represent the distance he has covered after n hours?*<br><br>a. 23, 33, 3+5n b. 20, 31, 5n c. 21, 33, 5n +1 d. 17, 27, 5n -3 |

## Data Analysis

The data was treated using the indices of both the CTT and the IRT models. Using the CTT model, a descriptive analysis was initially conducted by obtaining the mean (M) and standard deviation (SD) of test scores. Internal consistency (reliability) was computed using the *Cronbach's alpha*. Split-half reliability (odd-even) was also calculated in order to determine whether the scores within the same test are internally consistent or homogenous. These analyses were calculated using STATISTICA 9.0 software.

Using the IRT framework, the Rasch model (1PL) was employed. Important indices of psychometric properties of the instrument were calculated using the WINSTEPS ver. 3.60.1.9 software. The IRT indices include the unidimensionality of the test, hierarchical ordering of items, person reliability and separation estimates, as well as the

item reliability and separation estimates. Satisfying the fundamental assumptions of IRT, important indices were obtained and analyzed: (1) *Assumption of unidimensionality*: This was calculated using the coefficient of the ratio of the person separation reliability estimate of standard error to real error. According to Wright (1994) the person reliability estimated using standard error treats misfit as a random variation, while the person reliability estimated using real errors considers misfit as true departure from the unidimensionality of the construct. Considering certain provisions, the unidimensionality of the construct can be established if the coefficient yielded is closer to 1.0. (2.) The FIT index: the IRT assumption as regard to the goodness of fit is reflective of a match or fitting of the person's (examinee) ability with the difficulty of the items. Simply put, a person of a high ability is expected to respond correctly on items requiring high ability (particularly difficult items), while a person of a low ability is not expected to answer such items correctly. (3.) Reliability indices: Two indices of reliability were estimated using the WINSTEP software; one of which is Person Reliability and another is Item Reliability. (4.) Item Difficulty: this was estimated by identifying easy and difficult items. The bases of difficulty are the item characteristic curves (ICC's) generated in the 1PL –Rasch Model using WINSTEP software.

## Results

The descriptive analysis of test scores derived from the data of participants (N=180) provided the important psychometric indices of the developed instrument (*College Algebra Diagnostic Test)*. The mean of the total score was 9.67 and the standard deviation was 2.48. The overall reliability value in terms of coefficient alpha was .587. When split-half reliability (odd-even) was calculated it yielded a coefficient of .649.

When the test scores were analyzed using the 1PL-Rasch Model, new indices for reliability were obtained. The person reliability was .55 with RMSE of .72, and the item reliability was .97 with RMSE of .22. The errors (RMSE) in these estimates suggest how fitting the data are as regards the expected ability of test-takers and test difficulty. The error associated with the person reliability estimate is relatively high which indicates that the data may not adequately fit the expected ability of test-taker and the level of test difficulty. This was perhaps due to the occurrence of guessing the correct answer from the given choices; possibly how some low ability test-takers were able to answer difficult questions correctly. Item difficulty estimates were obtained using the ICC of each item in the test. The analysis of item measure provided an estimate of easy and difficult items in the test: *easy items*- item 1, 10, 9, 15, 13, 14, 11, 4, 2; *difficult items*- item 5, 8, 12, 7, 6, 3. The data revealed that item no. 5 is the most difficult item in the test yielding a parameter estimate of 1.96 and a model standard error of .18. On the other hand, item no.2 is the easiest item in the test yielding a parameter estimate of -2.89 and model standard error of .42 (see table 2.0).

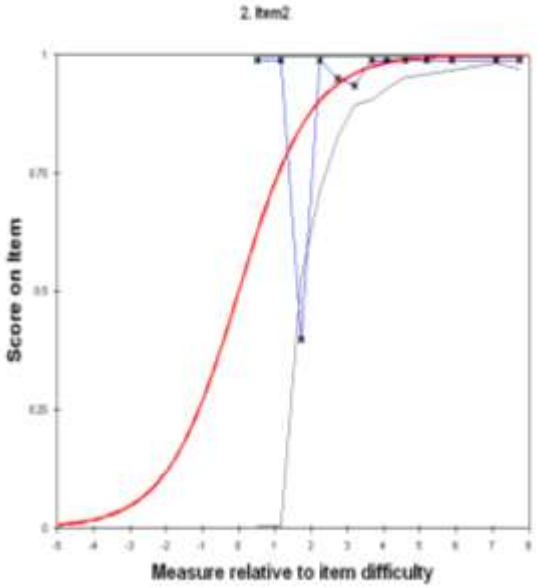



Figure 2Item No.2 level of difficulty

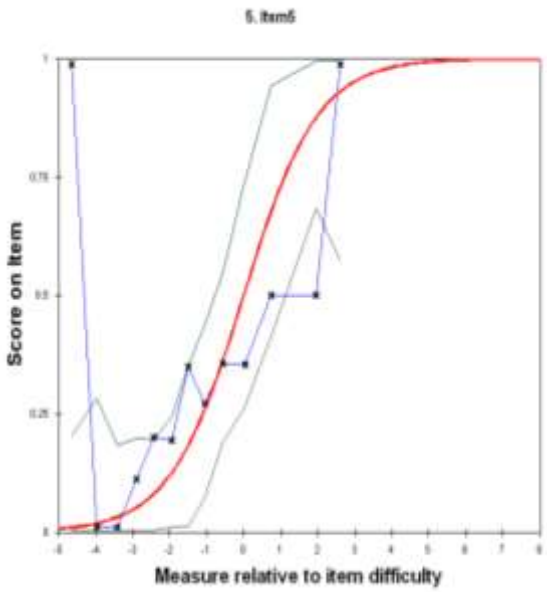



Figure 3.0 Item No. 5 level of difficulty

Table 2
*Item Statistic: Measure Order*

| ENTRY NUMBER | TOTAL SCORE | COUNT | MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PT-MEASURE CORR. | PT-MEASURE EXP. | EXACT MATCH OBS% | EXACT MATCH EXP% | ITEM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 55 | 180 | 1.96 | .18 | 1.24 | 2.6 | 1.50 | 3.2 | .26 | .46 | 70.2 | 75.7 | Item5 |
| 8 | 57 | 180 | 1.89 | .18 | 1.05 | .6 | 1.09 | .7 | .41 | .46 | 75.8 | 75.0 | Item8 |
| 12 | 63 | 180 | 1.70 | .18 | 1.07 | .9 | 1.11 | 1.0 | .40 | .46 | 71.9 | 73.3 | Item12 |
| 7 | 81 | 180 | 1.17 | .17 | 1.03 | .5 | 1.03 | .3 | .43 | .45 | 69.7 | 69.6 | Item7 |
| 6 | 93 | 180 | .84 | .17 | .92 | -1.4 | .89 | -1.2 | .51 | .44 | 68.5 | 68.7 | Item6 |
| 3 | 96 | 180 | .75 | .17 | 1.00 | .0 | .96 | -.4 | .44 | .44 | 68.0 | 68.6 | Item3 |
| 1 | 128 | 180 | -.18 | .18 | 1.17 | 2.1 | 1.16 | 1.0 | .25 | .38 | 69.7 | 73.3 | Item1 |
| 10 | 128 | 180 | -.18 | .18 | .94 | -.8 | .86 | -.9 | .43 | .38 | 75.3 | 73.3 | Item10 |
| 9 | 134 | 180 | -.38 | .18 | .98 | -.2 | .96 | -.1 | .37 | .36 | 75.3 | 75.8 | Item9 |
| 15 | 136 | 180 | -.45 | .19 | .88 | -1.4 | .70 | -1.8 | .47 | .35 | 76.4 | 76.7 | Item15 |
| 13 | 137 | 180 | -.48 | .19 | .90 | -1.2 | .71 | -1.7 | .46 | .35 | 79.2 | 77.1 | Item13 |
| 14 | 142 | 180 | -.67 | .20 | .99 | -.1 | .92 | -.3 | .35 | .33 | 78.1 | 79.5 | Item14 |
| 11 | 147 | 180 | -.87 | .21 | 1.01 | .1 | .93 | -.2 | .32 | .32 | 80.9 | 81.9 | Item11 |
| 4 | 169 | 180 | -2.22 | .32 | .89 | -.4 | .50 | -1.1 | .32 | .20 | 93.8 | 93.8 | Item4 |
| 2 | 174 | 180 | -2.89 | .42 | .96 | .0 | .49 | -.8 | .22 | .15 | 96.6 | 96.6 | Item2 |
| MEAN | 116.0 | 180.0 | .00 | .21 | 1.00 | .1 | .92 | -.2 | | | 76.6 | 77.3 | |
| S.D. | 37.8 | .0 | 1.37 | .07 | .10 | 1.1 | .25 | 1.2 | | | 8.3 | 7.9 | |

*Note.* WINSTEPS ver. 3.60.1.9 Output Table

Table 2.0 displays the *measure order* of items in the test. *Measure* refers to the estimate for the parameter or item difficulty. The difficulty of an item is defined to be the point on the latent variable at which its high and low categories are equally probable (Linacre, 2009). It is arranged according to the most difficult to the easiest item based on the reported logit value of each item in the test. In connection with the parameter estimates of each item measure, the standard error (SE) for each item is also reported. According to Linacre (2009, p. 27) "A standard error quantifies the precision of a measure or an

estimate. It is the standard deviation of an imagined error distribution representing the possible distribution of observed values around their "true" theoretical value".

In the analysis of goodness of fit, the average INFIT was 1.00 and the average OUTFIT was .92. This indicates that the data for the items showed goodness of fit satisfying the condition that the values should not exceed 1.50. To satisfy the assumption of unidimensionality, an indicative coefficient of unidimensionality was calculated by obtaining the ratio of the person separation reliability estimate using model standard error (.79) to the person separation reliability estimate using real standard error (.72). The total coefficient was 1.09; this highly indicates unidimensionality of the construct measured in the test considering that the value is close to unidimensionality index of 1.0.

The results definitively support the psychometric worthiness of the instrument (College Algebra Diagnostic Test). All the items passed the criteria of fit in the 1PL-Rasch Model, which indicates that the items in the test represent the expected ability of the target examinees. The data also supports the assumption of unidimensionality of the construct being measured by the items in the test, which is the pre-entry level of ability of test-takers in College Algebra.

The 1PL-IRT analyses provided a moderate and high level of reliability estimate of the instrument when data was tested for person reliability and item reliability respectively. This suggests that the instrument can reliably serve its purpose of diagnosing the level of ability of test-takers (low ability = novice; high ability = expert). However, the person reliability has barely yielded moderate reliability measure due to a high error estimate. The error estimate can be attributed to the chances of guessing the correct answer, which is the usual setback of ability tests with multiple choice response formats. Thus, a subsequent analysis of the effectiveness of distracters (wrong choices) is suggested in further calibration of the test; such analysis can be performed using 3PL-IRT. The analysis of data also provided an important aspect of the instrument's validity in terms of the goodness of fit. This indicates that the matching of item difficulty to person's ability (or vice-versa) was well-defined and corresponding.

## Discussion

Overall, it can be surmised that the proposed measure (College Algebra Diagnostic Test) possesses sound psychometric qualities that can be used to diagnose the level of expertise of students prior to pursuing the targeted subject domain. In light of the inherent value of correctly diagnosing academic ability early on, the said measure may allow educators and researchers to effectively address any potential pedagogical pitfalls that could encumber student learning.

It goes without saying that the need to conduct diagnostics of students' ability in a specific domain is of great importance. Through this, instructors can determine the existing levels of ability of learners that may hinder or facilitate subsequent learning. The results of diagnostics can also be useful in terms of instructional decision making; these can be a basis for educators in applying or implementing appropriate instructional interventions that would help maximize the academic improvement of learners. Such a benefit would in effect redound positively not just on the immediate academic concern of the student, but also potentially play a part in increasing learning in their other courses. In some academic institutions, diagnostic testing is used to identify students who are at risk of failure. Students

who are diagnosed to be at risk due to low ability are recommended for remediation or support programs (Coutis, Cuthbert, & MacGillivray, 2002). This offers a two-fold advantage: (1) it allows students diagnosed to be at risk for low ability to be given special attention, either by way of more focused instruction or through customized learning materials, (2) it also provides educators and school administrators with the potential to identify specific factors that are common to low-ability learners and address them accordingly.

With this, diagnostics as an academic practice should not be viewed as only focusing on cognitive aspects, but also one that may encompass other factors that have significant impact on student's performance. Consider for instance the case of a student who has very low confidence in mathematics; such an affective condition can be debilitating as explained by Dweck (1986), wherein the students' level of confidence (combined with his motivation) determines his behavior pattern and the level of persistence he will employ in performing mathematical tasks. Thus, failure to diagnose students' learning ability, traits, and experience may lead to significant shortcomings in an institution's obligation to effectively educate. It is clearly evident that diagnostics in this example would serve to benefit both the learner and the educator not only by identifying a low-ability area, but also a factor or group of factors that affect it in a negative manner. While it is true that a unilateral diagnostic (single aspect) of mathematical ability of learners is not enough to capture the actual ability and potential of students to contribute to their success in the subject domain, some universities and academic institutions are concerned on the use of multilateral diagnostics. The many-sided screening can be composed of a number of good predictors to academic success in mathematics such as: the foundation (prior) knowledge, motivational aspects, and aptitude (Carmody, Godfrey, & Wood, 2006).

Ultimately, the intention of designing and constructing diagnostic tests to assess the ability of students in mathematics courses like Algebra can be more worthy if the test is psychometrically calibrated using a more advanced and reliable test analysis model such as the IRT. This model of test analysis provides a more practical and scientific way of ensuring the reliability and validity of the test. This will then ultimately result in a more accurate measure of the desired criterion as explained in this study.

Certain recommendations can thus be inferred based on the results of the present study to further establish the psychometric properties of the designed instrument; (1) the use of more advanced IRT models (2PL or 3PL) is preferred, particularly in addressing the issues on the probability of guessing and a more thorough analysis of item discrimination, (2) a larger number of samples with varying year levels (secondary and tertiary) should be included in order to increase the probability of including experts and novices in the domain, (3) the use of other techniques (post-interview, think aloud, etc.) can be used as part of the diagnostic procedure to further analyze the test-takers' experience during the conduct of the test, and (4) the inclusion of affective and motivational aspect as part of the diagnostics items. Taken together, these measures will help ensure that the results of future similar lines of study will employ only the most rigorous scientific criterion possible.

## References


Baroody, A., & Ginsburg, H. (1993). The effects of instruction on children's understanding of the "equals" sign. *Elementary School Journal, 84*, 199–212.

Carmody, G., Godfrey, S., & Wood, L. (2006). *Diagnostic tests in a first year mathematics subject.* UniServe Science 2005 Conference Proceedings [online site]. Retrieved from http://science.uniserve.edu.au/pubs/procs/2006/carmody.pdf

Coutis, P., Cuthbert, R., & MacGillivray, H. (2002) *Bridging the gap between assumed knowledge and reality: a case for supplementary learning support programs in tertiary mathematics.* Proceedings of Engineering Mathematics and Applications Conference, The Institution of Engineers, Australia [online site]. Retrieved from http://eprints.qut.edu.au/25756/.

Dweck, C. (1986). Motivational process affecting learning. *American Psychologist, 41*, 1040-1048.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: issues and practice, 12*(3), 535-556.

Hambleton, R. K., & Rogers, J. H. (1990). Using item response models in educational assessments. In W. Schreiber & K. Ingenkamp (Eds.), *International developments in large-scale assessment* (pp. 155-184). England: NFER-Nelson.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19,* 509–539.

Kaplan, R. M. & Saccuzo, D. P. (1997). *Psychological testing: Principles, applications and issues*. Pacific Grove: Brooks Cole Pub. Company.

Ketterlin-Geller, L. R., & Yovanoff, P. (2009). Diagnostic Assessments in Mathematics to Support Instructional Decision Making. *Practical Assessment, Research and Evaluation* [online site]. Retrieved from http://pareonline.net/pdf/v14n16.pdf

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education, 37*, 297-312.

Lerch, C. M. (2004). Control decisions and personal beliefs: Their effect on solving mathematical problems. *Journal of Mathematical Behavior, 23*, 21-36.

Linacre, J. M. (2009). A User's Guide to WINSTEPS: MINISTEPS Rasch Model Computer Programs. *Program Manual 3.69.0*. winsteps.com

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Magno, C. P., & Oano, J. A (2010). *Designing written assessment for student learning*. Manila: Phoenix Publishing House, Inc.

Payne, D. A (1992). *Measuring and evaluating educational outcomes*. NY: McMillan Publishing Company.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Ramos, ML. F. (2008, August). *Construction and Evaluation of a Diagnostic Examination in College Algebra for Freshmen of the College of Science, University of Santo Tomas.* Proceedings at the Mathematics Teachers Association of the Philippines Tertiary Level Annual Convention [online site]. Retrieved from http://www.eric.ed.gov/PDFS/ED510031.pdf

Rittle-Johnson, B., Star, J. R., & Durkin K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology, 101*(4), 836–852.

Rittle-Johnson, B., & Kmicikewycz, A. O. (2008). When generating answers benefits arithmetic skill: The importance of prior knowledge. *Journal of Experimental Child Psychology, 101,* 75–81.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. New York: Academic Press.

Schoenfeld, A. H., & Hermann, D. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8,* 484 – 494.

Sebrechts, M., Enright, M., Bennett, R., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction*, *14*(3), 285–343.

Wiberg, M. (2004). Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving license-test. *EM No. 50, 2004, ISSN 1103-2685.*

Wright, B. D. (1994). A Rasch unidimensionality coefficient. *Rasch Measurement Transactions, 8*(3), 385.

## About the Authors

**Jonathan V. Macayan** is a graduate of Bachelor of Science in Psychology with Master's degree in the same field. He has completed the academic requirements for PhD. in Psychology in De La Salle University specializing in Quantitative Analysis. As a practicing psychologist, he has exposures on the diverse fields of psychological profession including 2 years of Industrial-Organizational experience, 12 years of college teaching and 6 years of educational administration covering 1 year of holding the position as Area Coordinator for the College of Arts and Sciences in AMACC and 5 years of managing the MIT Department of Psychology as the Program Head/Chairperson. He is also active in the practice of psychological profession on the following areas: Psychological Research consultancy and Organizational development & Soft skills Training consultancy. At present, Prof. Macayan is a faculty member of SLHS-Psychology in MIT and concurrently holding the position of department chairperson for Psychology department.

For correspondence you may contact Prof. Macayan in Tel. No: 247-5000, loc.1411, Cell No: 0908-7518536 Email: jomacayan2001@yahoo.com; jvmacayan@mapua.edu.ph

**Bernardino C. Ofalia** is a graduate of Bachelor of Arts in Psychology with Master's degree in Industrial Psychology. He has also earned his degree of Doctor of Education major in Educational Administration from Pamantasan ng Lungsod ng Maynila (University of the City of Manila). He has been in the academe for more than 15 years holding the following

positions: Associate Professor (College of Human Development) and Program Director (Guidance Counseling Office). He conducts lecture on the fields of Industrial and Applied Psychology both in the undergraduate and graduate level. His exposures are so varied in the area of career assistance, placement, test development, and group and individual counseling. At present he is a fulltime faculty of the Department of Psychology at PLM.

For correspondence you may contact Prof. Ofalia in Tel. No: 799-8814
Cell No: 0906-5119010; Email: bernie_ofalia@yahoo.com

Educational Measurement and Evaluation Review (2011), Vol. 2, 77-91


http://pemea.club.officelive.com/ EMEReview.aspx

# Exploring the Factors of Perfectionism within the Big Five Personality Model among Filipino College Students

**Joel C. Navarez**
*De La Salle University – Manila*

**Ryan Francis O. Cayubit**
*De La Salle University – Manila*
*University of Santo Tomas*

**Abstract** The study dealt with determining if perfectionism through its different scales would have an effect on the five factors of personality. The scales of perfectionism include Concern over Mistakes, Personal Standards, Parental Expectations, Parental Criticism, Doubt about actions and Organization. The five factors of personality tested include Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness. The study aimed to better understand the complex set of behaviors and characteristics associated with perfectionism. The study is based on the theory of Perfectionism by Frost and the Big Five Factor Theory of Personality. It is a cross-sectional predictive investigation that made use of SEM with bootstrapping on data gathered from 106 college students. The results revealed that the proposed model showed adequate goodness of fit, as indicated by the low chi-square and discrepancy function values ($\chi$2=1180.258, *df*=589, $\chi$2/*df*=2.004) which is significant at .05 alpha level. The sample represents the model well as shown by acceptable fit indices (*CFI* =.531, *PCFI* =.496, *RMSEA* =.098).

**Key words:** *Perfectionism, Big Five Personality, Adolescents, Structural Equations Modeling*

## Introduction

Over the past two decades, research on the construct of perfectionism has made great progress in understanding its nature, correlates, and consequences. Recent investigations have stressed that perfectionism is a complex and multidimensional construct (Frost, Marten, Lahart, & Rosenblate, 1990; Hewitt & Flett, 1991; Hewit, Flett, Besser, Sherry, & McGee, 2003), which is manifested in different ways and has links to various adaptive and maladaptive behaviors (Enns &Cox, 2002).

Frost et al. (1990) emphasized that perfectionism is the tendency for overly critical evaluation of one's own behavior, expressed in over concern for mistakes and uncertainty regarding actions and beliefs. Moreover, Frost and co-authors have pointed out that

perfectionists place considerable value on their parents' expectations and evaluations. They have also been described to overemphasize order, organization, and neatness. In the Frost Multidimensional Perfectionism Scale (FMPS), Frost, Marten, Lahart, and Rosenblate (1990) stated that perfectionism has the following six (6) subscales: Concern over Mistakes (CM), Personal Standards (PS), Parental Expectations (PE), Parental Criticism (PC), Doubts about actions (D), and Organization (O). Whereas the first five scales represent the core perfectionism dimensions, organization was found to be only loosely related to other scales.

Perfectionism has a long history in both clinical and personality research. Empirical studies have demonstrated perfectionism as a complex phenomenon closely linked with normal psychological functioning, as well as emotional and behavioral difficulties (Adler, 1956; Beck, 1967; Blatt, 1995; Chang, Watkins, & Banks, 2004; Shafran & Mansell, 2001). Scores on the FMPS have been related to a variety of problems such as competition anxiety in athletes (Frost & Henderson, 1991), evaluation anxiety in college students (Frost & Marten, 1990), insomnia (Lundh, Broman, Hetta, & Saboonchi, 1994), social phobia (Juster, Heimberg, Frost, Holt, Mattia, & Faccenda, 1996), obsessive-compulsive symptoms (Rhéaume, Freeston, Dugas, Letarte, & Ladouceur, 1995), eating disorders like anorexia nervosa (Bardone-Cone, Wonderlich, Frost, Bulik, Mitchell, & Uppalla, 2007; Bastiani, Rao, Weltzin, & Kaye, 1995; Bastiani, Rao, Weltzin, & Kaye, 1995; Halmi, Sunday, Strober, Kaplan, Woodside, & Fichter, 2000; Srinivasagam, Kaye, Plotnicov, Greeno, Weltzin, & Rao, 1995) suicidal preoccupation (Adkins & Parker, 1996), and psychological distress (Frost, Marten, Lahart, & Rosenblate, 1990; Hewitt & Flett, 1991b; Hewitt, Newton, Flett, & Callander, 1997; Enns, Cox, Sareen, & Freeman, 2001; Kawamura, Hunt, Frost, & DiBartolo, 2001; DiBartolo, Frost, Chang, LaSota, & Grills, 2004; Dunkley, Blankstein, Masheb, & Grilo, 2006). In all these studies, problematic outcomes were most closely related to the FMPS subscales assessing evaluation concerns (CM, PE, PC, and D). In contrast, the other two subscales (PS and O) have shown relations with more desirable outcomes such as success orientation (Frost & Henderson, 1991), achievement motivation (Adkins & Parker, 1996), and goal commitment (Flett, Sawatzky, & Hewitt, 1995).

Another central question of the research literature on perfectionism which is not given much focus is isolating how each of the dimensions of perfectionism influences important personality constructs. Today's most prevalent system to describe personality traits is the "Big Five" personality system in which personality can be described by five broad trait dimensions: neuroticism, extraversion, openness, agreeableness, and conscientiousness (Costa & McCrae, 1992; John & Srivastava, 1999). It is the general goal of this study to locate the extracted factors of perfectionism within a comprehensive scheme of personality–the five-factor model (FFM)—in order to gain a better understanding of what these higher-order factors represent. Literature reviews have suggested that the FFM is a useful heuristic framework relevant to the description and understanding of specific vulnerability styles (e. g., Widiger & Costa, 2002).

One investigation (Flett, Hewitt & Dyck, 1989) found perfectionism to be associated with neuroticism as measured by the Eysenck Personality Inventory (Eysenck & Eysenck, 1968). In an effort to explain this correlation, perfectionism was described to be associated with a fear of negative evaluation, desire for social approval, and indices of emotional instability such as depression and anxiety (Hewitt, Flett, & Blankstein, 1991).

Several of the dimensions of perfectionism described by Frost and colleagues (1990), especially high personal standards and a preference for order and organization

were associated with the Big Five factor of conscientiousness specifically with the traits of good work habits and striving for high achievement. The positive achievement striving factor appears to capture an adaptive aspect of perfectionism that involves painstaking effort, high personal standards, a need to excel, and good organization among other qualities. These characteristics appear similar to the Big Five factor of conscientiousness which includes the following facets: competence, order, dutifulness, achievement striving, self-discipline, and deliberation.

When considering the Big Five factor of extraversion and openness, only minimal association with perfectionism may be anticipated. Individuals high in perfectionism might be expected to be less open to new ideas, actions and values as they fear failure and thus might prefer the tried and true familiar experiences to novel experiences and the uncertainty involved. Individuals high in perfectionism might be expected to score low in agreeableness as they have demanding and unrealistic expectations of others.

Although there has been a growing interest in the construct of perfectionism, compared with the adult literature, relatively few studies have involved investigation on perfectionism and personality among adolescent samples (O'Connor, 2007; Rice & Preusser, 2002). Recognizing that perfectionism is prevalent among college students, for whom academic performance is crucial to one's personal development and also considering that they are in a period where personality and identity are formed, exploring the construct of perfectionism and personality characteristics would be highly essential.

There were some empirical studies that pointed out the need to see perfectionism as a complex phenomenon in college. Castro and Rice (2003) have found that perfectionism significantly predicted the self-reported academic achievement (assessed by Grade Point Average) among Asian and African American students, with some characteristics beneficial and others an impediment. Chang et al. (2004) assessed racial variation of white and black college females in how adaptive and maladaptive perfectionism related to psychological functioning, and found only the latter, and not both, associated with stress.

In relation to the previous discussions, the first goal of the current study was to examine the association between dimensions of perfectionism and the big five personality model. Flett, Hewitt, Oliver, and Macdonald (2002) postulated that dimensions of perfectionism might be associated with a contingent sense of self-worth and low unconditional self-acceptance. Similarly, Ellis (2002) pointed to the general hyper competitiveness of perfectionists and suggested that they may exhibit a personality style that is focused excessively on evaluative outcomes, suggesting a link between perfectionism and a diminished sense of unconditional self-acceptance. The present study tested one model using Structural Equations Modeling (SEM) that focuses on the six factors of the construct as it affects personality. In the model, it is hypothesized that the six different perfectionism factors when taken together contributes to increased variance in personality in relation to the 30 facets of the five-factor model. The model would further describe how each of the personality constructs is influenced by the dimensions of perfectionism.

# Method

## Research Design

Based on the research classification of Johnson (1991), the present study is cross-sectional and predictive in nature. It is cross-sectional in nature because the data collected from the participants cover a brief period of time. Likewise, the data collected directly apply to the participants and comparisons are made across the variables of interest namely: Perfectionism and personality. Moreover, the present endeavour is also predictive in nature as the main focus is to forecast one's personality based on his perfectionistic tendencies in the absence of experimental manipulation (Johnson, 1991).

## Participants

A total of 106 college students participated in the study. They were selected using non-probability sampling specifically convenience sampling. Only college students who were interested and available at the time of test administration were included in the sample. The participants are second year and third year psychology students of the University of Santo Tomas enrolled in the course Behavior Measurement and Analysis and Experimental Psychology respectively.

## Research Instrument

**Measures of Perfectionism.** The Multidimensional Perfectionism Scale is a 35 item questionnaire designed to measure perfectionism. It has 5 subscales that are answered on a 5-point scale. They are Concern over Mistakes (CM), Personal Standards (PS), Parental Expectations (PE), Parental Criticism (PC), Doubts about Actions (D), and Organization (O). The instrument was developed by Frost, Lahart and Rosenblate in 1991. Internal consistency information of the instruments are as follows: .90 for the overall perfectionism, .88 for concern over mistakes, .83 for personal standards, .84 for parental expectations, .84 for parental criticism, .77 for doubts about actions, and .93 for organization. Validity evidences of the scale is based on the convergent method wherein high correlation with other measures of perfectionism was found, specifically the Burns' Perfectionism Scale (Burns, 1980), the Self-Evaluative (SE) Scale from the IBT (Jones, 1968), the Perfectionism Scale from the EDI (Garner et al., 1983), and the Self-Oriented Perfectionism and Socially-Prescribed Perfectionism scales on Hewitt and Flett's (1991) Multidimensional Perfectionism Scale (Frost et al., 1991).

**Revised NEO Personality Inventory.** The Revised NEO Personality Inventory or the NEO PI-R was used to measure the different factors of the personality of the college students. It is a concise measure of the five major dimensions of personality and some important traits that define each domain. It consists of 240 items answered on a 5-point scale. It is self-administered for men and women of all ages. The instrument measures the facets of Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A), and Conscientiousness (C). Reliability of the instrument was determined through internal consistency and test-retest method. Reliability coefficient for the scales ranged from .66 to .92. In addition, a combination of rational and factor analytic method was used to establish

the validity of the instrument. Based on varimax-rotated principal components, correlations between the factor scores and the N, E, O, A, and C domain scales were .91, .89, .95, .95, and .89, respectively (Costa & McCrae, 1992).

## Procedure

Test administration was done in groups during the experimental psychology class of the third year participants and in the behaviour measurement and analysis class of the second year participants. They were oriented that they would be asked to answer two questionnaires that would help them to get to know aspects about themselves better. The first test that was administered was the MPS. On the average, test administration for the instrument was around 30 minutes. This was followed by the administration of the NEO-PIR which took the participants around 1 hour and a half. During the administration, standard procedures and test instructions were followed based on the manual of both instruments. The instruments were then scored and interpreted and subjected to data analysis in order to answer the objectives of the study.

## Data Analysis

Data gathered were analyzed using the Statistical Package for Social Sciences version 16 and the Analysis of Moment Structure version 16. Descriptive statistics specifically the M and SD were used to determine the levels of perfectionism and personality of the participants. In addition, SEM analysis was used to test a model showing perfectionism predicting personality. The model is tested for goodness of fit using the chi-square ($\chi 2$), Root Mean Square Error Approximation (*RMSEA*), Comparative Fit Index (*CFI*), and Population Comparative Fit Index (*PCFI*)

# Results

In the analysis, the means and standard deviations are reported for each factor. Correlations were conducted among the factors of personality and perfectionism, the, SEM model was conducted to test the effect of perfectionism on personality. Table 1 shows the mean scores, standard deviation, and Cronbach's alpha for NEO-PI-R and FMPS.

In assessing the students' personality using the NEO-PI-R, the M and SD obtained are as follows: High in neuroticism (*M*=106.21, *SD*=15.75), average in extraversion (*M*=117.67, *SD*=17.00), high in openness to experience (*M*=117.99, *SD*=14.69), low in agreeableness (*M*=110.45, *SD*=14.62), and low in conscientiousness (*M*=107.64, *SD*=18.05). Results show that the participants are prone to experience anxiety, anger and hostile feelings. Likewise, when faced with stressful situations, they appear to have difficulty handling it. Though they may be characterized as outgoing, curious and sociable, they appear to be always on guard. They are often skeptical and may question the motives and reasons of others. In addition, they are less helpful and altruistic particularly if they will not get anything in return. Nonetheless, they appear to be open to new ideas and ways of doing things and are often not afraid to venture out into the unknown.

Table 1
*Means and Standard Deviations of Dimensions of Five Personality Factor and Perfectionism*

| | *M* | *SD* | Cronbach's Alpha |
|---|---|---|---|
| Neuroticism | 106.20 | 15.75 | .806 |
| Extraversion | 117.67 | 17.00 | .79 |
| Openness to Experience | 117.99 | 14.69 | .80 |
| Agreeableness | 110.45 | 14.62 | .81 |
| Conscientiousness | 107.64 | 18.05 | .79 |
| Concern over Mistakes | 22.71 | 5.89 | .77 |
| Personal Standards | 22.04 | 4.02 | .75 |
| Parental Expectations | 15.85 | 3.79 | .76 |
| Parental Criticism | 9.54 | 2.95 | .77 |
| Doubts about actions | 11.84 | 3.23 | .78 |
| Organization | 21.53 | 3.69 | .79 |

Concerning perfectionism, the following descriptive statistics were obtained; Concern over Mistakes (*M*=22.71, *SD*=5.89), Personal Standards (*M*=22.04, *SD*=4.02), Parental Expectations (*M*=15.85, *SD*=3.79), Parental Criticism (*M*=9.54, *SD*=2.95), Doubts over actions (*M*=11.84, *SD*=3.23), and Organization (*M*=21.53, *SD*=3.69). This is an indication of the tendency of the participants to exhibit behaviour in relation to the scales of MPS. Acceptable internal consistencies as shown by Cronbach's alpha of the items were obtained for all measures.

To establish the relationship among the variables involved in the model, a zero-order correlation was conducted to determine the pair of variables that are significantly related (Table 2). Doubt about action is positively correlated with neuroticism, personal standard is positively correlated to extraversion, concern over mistakes and parental criticism is negatively correlated to agreeableness, and personal standard is positively correlated to conscientiousness. The aforementioned correlations are all significant at .05 level of significance.

Table 2
*Zero-order correlation of the factors of personality and perfectionism*

| | Perfectionism | | | | | |
|---|---|---|---|---|---|---|
| | Concern over Mistakes | Personal Standards | Parental Expectations | Parental Criticism | Doubts about actions | Organization |
| Neuroticism | 0.09 | -0.03 | 0.04 | 0.02 | **0.21*** | -0.01 |
| Extraversion | -0.10 | **0.25*** | 0.15 | -0.02 | -0.06 | 0.14 |
| Openness to Experience | -0.15 | 0.04 | 0.19 | 0.04 | -0.07 | 0.07 |
| Agreeableness | **-0.19*** | -0.05 | -0.12 | **-0.23*** | -0.07 | 0.07 |
| Conscientiousness | -0.02 | **0.38*** | 0.08 | 0.05 | -0.09 | **0.30*** |

*$p<.05$

The estimates of regression weight were computed in order to determine the effect of perfectionism on personality factors (Table 3). When perfectionism goes up by 1, neuroticism goes up by 0.5. extraversion increases by 0.39, openness to experience is

lowered by 0.082, agreeableness decreases by 0.951, and conscientiousness goes up by 0.919. Results however indicated that the regression weights for perfectionism in the prediction of personality factors is not significantly different from zero at the 0.05 level (two-tailed).

Table 3
*Estimates of Regression Weight of the Big Five Personality to Perfectionism*

| | Parameter Estimate | Standard error | Critical Ratio | p-level |
|---|---|---|---|---|
| Neuroticism | .590 | 1.053 | .561 | .575 |
| Extraversion | .394 | .989 | .398 | .691 |
| Openness to Experience | -.082 | .745 | -.110 | .912 |
| Agreeableness | -.951 | 1.106 | -.860 | .390 |
| Conscientiousness | .919 | 1.182 | .777 | .437 |

The model showed adequate goodness of fit, as indicated by the low chi-square and discrepancy function values ($\chi^2$=1180.258, *df*=589, $\chi^2/df$=2.004). The minimum chi-square value was significant at .05 alpha level showing that the departure of the data from the model is significant at the .05 level. The *CFI* (.531) and *PCFI* (.496) and the *RMSEA* (.098) indicated an acceptable fit. These indicate that the sample represents the model well.

## Discussion

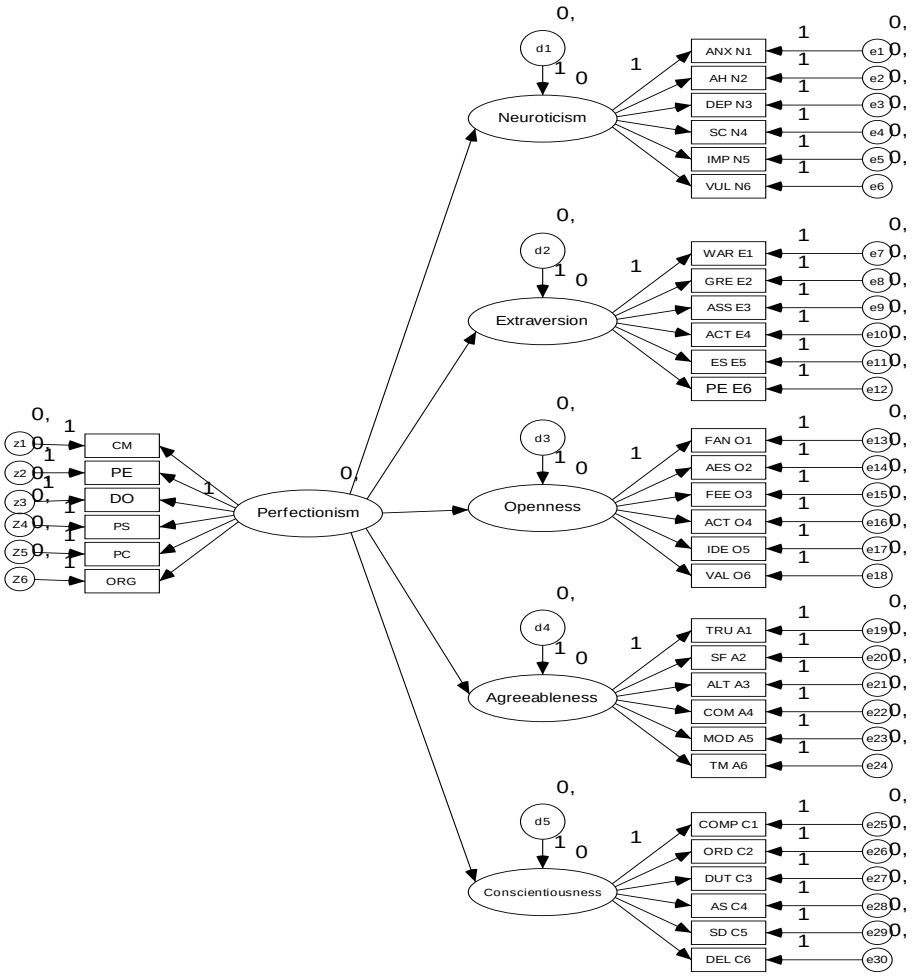
The results of the present study support previous investigations that have highlighted a dichotomy on the personality structure of a perfectionist, which has been labelled as 'normal or neurotic' (Hamachek, 1978; Hollender, 1965), 'adaptive or maladaptive' (Cox, Enns, & Clara, 2002), 'positive or negative' (Terry-Short, Owens, Slade, and Dewey, 1995), or 'healthy or unhealthy' (Stumpf & Parker, 2000). The functional aspect of perfectionism is characterized by setting goals and striving for rewards, while maintaining flexibility and satisfaction with self. Conversely, the dysfunctional aspect of perfectionism has been described as setting rigid goals, high standards, an inability to feel a sense of fulfilment, and distress over one's capability (Enns & Cox, 2005).

When perfectionism factors were correlated with personality dimensions, the association between doubts about actions and neuroticism was established, which implies that maladaptive perfectionists have doubts about the quality of their performance as manifested in their difficulty adapting to new situations or conditions (Goldberg, 2001), self-consciousness, impulsivity, and vulnerability similar to the facets of neuroticism dimension as described by Costa and McCrae (1992b).

Another significant finding was that maladaptive perfectionists' concern over mistakes which are accompanied by tendencies for overly critical evaluations of one's own behavior" (Frost et al. 1990), and their overall perception of frequent inability to meet parental standards negatively influence the agreeableness dimension of their personality which is associated with being courteous, flexible, trusting, cooperative, tolerant, and treating others fairly and kindly (Costa & McCrae, 1992b).

Figure 1

*Model one of the effects of perfectionism on personality.*




The adaptive factors of perfectionism showed an association with the extraversion and conscientiousness dimensions of the five-factor personality model. Personal standards of the perfectionism dimension located in the extraversion personality factor suggested that perfectionism can be interpreted as initiative, surgency, ambition, and impetuousness (Hogan, 1986). Meanwhile, the organization factor of perfectionism which emphasizes order and precision is linked to the conscientiousness personality factor which explains a perfectionist's tendency for high degree of dependability, organization, persistence, and achievement-orientation (Costa & McCrae, 1992a).

Having established the correlation of perfectionism factors with specific personality dimensions, there are specific character traits and behavior that may differentiate maladaptive from adaptive perfectionists. This present finding is relevant and related to researches done in the past (Bieling, Israeli, & Antony, 2004; Dunkley & Blankstein, 2000; Frost, Heimberg, Holt, Mattia, & Neubauer, 1993; Enns, Cox, Sareen, & Freeman, 2001; Terry-Short, Owens, Slade, & Dewey, 1995). The hypothesized model explaining

perfectionism as an integrated construct affecting the five personality factors was confirmed in this study. The study has established the model of perfectionism predicting variance in personality scores. It further highlighted the need to take personality traits into account when explaining why some people are happy even with less-than-perfect results as compared to those who strive for perfection and perceive imperfection as unacceptable.

However, there are areas of the present study that warrant attention in future research. Since findings were based on self-report measures, replication with other methods of data collection (e. g., diaries, observer ratings) would be beneficial. The generalizability of the present results should be examined in other student populations, different age groups, and clinical populations.

Since it has been established that perfectionism can predict personality factors, practical use and implication can be put forth. The present study has looked into the different structures of perfectionism. This could be a spring board for further studies particularly those related to assessment and evaluation. With the new models of perfectionism, additional instruments could be constructed. This is similar to the studies of Stober (1998) and Khawaja and Armstrong (2005) where they explored the structures of perfectionism and found out that the 6 original dimensions identified by Frost et al. (1991) no longer exist in German and Australian samples respectively.

Finally, the findings might also inform counselors working in the school setting to further encourage healthy/adaptive or address unhealthy/maladaptive perfectionistic thinking among students. According to Beck (1976), poor psychological adjustment is often considered a function of negative automatic thoughts that distort the way we interpret ourselves and the world around us. Consistent with this view, counselors or educational researchers might develop interventions to help students identify such automatic thoughts associated with the need to be perfect for others, examine these thoughts for distortions, and then restructure or eliminate these thoughts so that they no longer have harmful influences on their feelings and behaviors (e.g., Ferguson & Rodway, 1994).

## References

Adkins, K. K., & Parker, W. D. (1996). Perfectionism and suicidal preoccupation. *Journal of Personality*, *64*, 529-543.

Adler, A. (1956). The neurotic disposition. In H. L. Ansbacher & R. R. Ansbacher (Eds.), *The individual psychology of Alfred Adler*. New York: Harper.

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment, 10*, 176–181.

Antony, M. M., Purdon, C. L., Huta, V., & Swinson, R. P. (1998). Dimensions of perfectionism across the anxiety disorders. *Behaviour Research and Therapy, 36*, 1143-1154.

Bardone-Cone, A. M., Wonderlich, S. A., Frost, R. O., Bulik, C. M., Mitchell, J. E., Uppalla, S., et al. (2007). Perfectionism and eating disorders: Current status and future directions. *Clinical Psychology Review, 27*, 384–405.

Bastiani, A. M., Rao, R., Weltzin, T. E., & Kaye, W. H. (1995). Perfectionism in anorexia nervosa. *The International Journal of Eating Disorders, 17*, 147–152.

Beck, A. T. (1967). *Depression: Clinical, experimental and theoretical aspects*. New York: Harper and Row. Beck, A. T. (1976). *Cognitive therapy and emotional disorders*. New York: International Universities Press.

Beck, A. T., & Freeman, A. (1990). *Cognitive therapy of personality disorders*. New York: Guilford Press.

Bieling, P. J., Israeli, A. L., & Antony, M. M. (2004). Is perfectionism good, bad, or both? Examining models of the perfectionism construct. *Personality and Individual Difference, 36*, 1373–1385.

Bieling, P. J., Israeli, A., & Antony, M. M. (2003). Is perfectionism good, bad, or both? Examining analysis. *Behaviour Research and Therapy, 40*, 773–791.

Bieling, P. J., Israeli, A., Smith., J., & Antony, M. M. (2003). Making the grade: The behavioural consequences of perfectionism in the classroom. *Personality and Individual Differences, 35*, 163–178.

Blatt, S. J. (1995). The destructiveness of perfectionism: Implications for the treatment of depression. *American Psychologist, 50*(12), 1003–1020.

Blatt, S. J. (2004). *Experiences of depression: Theoretical, clinical, and research perspectives*. Washington, DC: American Psychological Association.

Blatt, S. J., & Shichman, S. (1983). Two primary configurations of psychopathology. *Psychoanalysis and Contemporary Thought, 6*, 187–254.

Brown, E. J., Makris, G. S., Juster, H. R., Leung, A. W., Heimberg, R. G., & Frost, R. O. (1999). Relationship of perfectionism to affect, expectations, attributions, and performance in the classroom. *Journal of Social and Clinical Psychology, 18*, 98–120.

Bulik, C. M., Tozzi, F., Anderson, C., Mazzeo, S. E., Aggen, S., & Sullivan, P. F. (2003). The relation between eating disorders and components of perfectionism. *The American Journal of Psychiatry, 160*, 366–368.

Burns, D. D. (1980). The perfectionist's script for self-defeat. *Psychology Today, November*, 34-51.

Burns, D. D. (1983). The spouse who is a perfectionist. *Medical Aspects of Human Sexuality, 17*, 219-230.

Castro, J. R., & Rice, K. G. (2003). Perfectionism and ethnicity: Implications for depressive symptoms and self-reported academic achievement. *Cultural Diversity and Ethnic Minority Psychology, 9*(1), 64–78.

Chang, E. C., Watkins, A. F., & Banks, K. H. (2004). How adaptive and maladaptive perfectionism related to positive and negative psychological functioning: Testing a stress-mediation model in black and white female college students. *Journal of Counseling Psychology, 51*(1), 93–102.

Clark, D. A., Steer, R. A., Beck, A. T., & Ross, L. (1995). Psychometric characteristics of revised sociotropy and autonomy scales in college students. *Behaviour Research and Therapy, 33*, 325–334.

Coles, M. E., Frost, R. O., Heimberg, R. G., & Rheaume, J. (2003). "Not just right experiences": Perfectionism, obsessive-compulsive features and general psychopathology. *Behaviour Research & Therapy, 41,* 681-700.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cox, B., J., Enns, M. W., & Clara, I. P. (2002). The multidimensional structure of perfectionism in clinically distressed and college student samples. *Psychological Assessment, 14*, 365-373.

DiBartolo, P. M., Frost, R. O., Chang, P., LaSota, M., & Grills, A. E. (2004). Shedding light on the relationship between personal standards and psychopathology: The case for contingent self-worth. *Journal of Rational-Emotive and Cognitive Behavior Therapy, 22*, 241–254.

Dunkley, D. M., Blankstein, K. R., & Flett, G. L. (1997). Specific cognitive-personality vulnerability styles in depression and the five-factor model of personality. *Personality and Individual Differences, 23,* 1041–1053.

Dunkley, D. M., Blankstein, K. R., Masheb, R. N., & Grilo, C. M. (2006). Personal standards and evaluative concerns dimensions of "clinical" perfectionism: A reply to Shafran et al. (2002, 2003) and Hewitt et al. (2003). *Behaviour Research and Therapy, 44*, 63–84.

Dunkley, D. M., Zuroff, D. C., & Blankstein, K. R. (2003). Self-critical perfectionism and daily affect: Dispositional and situational influences on stress and coping. *Journal of Personality and Social Psychology, 84*, 234–252.

Dunkley, D., Blankstein, K., Zuroff, D., Lecce, S. & Hui, D. (2006). Self-critical and personal standards factors of perfectionism located within the five-factor model of personality. *Personality and Individual Differences, 40*, 409–420.

Dunkley, D. M., Blankstein, K.R., Halsall, J., Williams, M., & Winkworth, G. (2000). The relation between perfectionism and distress: Hassles, coping, and perceived social support as mediators and moderators. *Journal of Counseling Psychology, 47*, 437–453.

Einstein, D. A., Lovibond, P. F., & Gaston, J. E. (2001). Relationship between perfectionism and emotional symptoms in an adolescent sample. *Australian Journal of Psychology, 52*, 89-93.

Ellis, A. (1962). *Reason and emotion in psychotherapy*. New York: Lyle Stuart.

Ellis, A. (1977). Psychotherapy and the value of a human being. In A. Ellis & R. Greiger (Eds.), *Handbook of rational-emotive therapy* (pp. 99–112). New York: Springer.

Ellis, A. (1995). Changing rational-emotive therapy (RET) to rational emotive behavior therapy (REBT). *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *13*, 85–89.

Ellis, A. (2002). The role of irrational beliefs in perfectionism. In G. L. Flett & P. L. Hewitt (Eds.), *Perfectionism: Theory, research, and treatment* (pp. 217–229). Washington, DC: American Psychological Association.

Enns, M. W., & Cox, B., J. (2005). The nature and assessment of perfectionism: A critical analysis. In G. L. Flett & P. L. Hewitt (Eds.), *Perfectionism: Theory, Research and Treatment.* Washington: American Psychological Association.

Enns, M. W., Cox, B. J., Sareen, J., & Freeman, P. (2001). Adaptive and maladaptive perfectionism in medical students: A longitudinal investigation. *Medical Education, 35*, 1034–1042.

Ferguson, K. L., & Rodway, M. R. (1994). Cognitive behavioural treatment of perfectionism: Initial evaluation studies. *Research on Social Work Practice*, 4, 283-308.

Flett, G. L., & Hewitt, P. L. (2002). Perfectionism and maladjustment: An overview of theoretical, definitional, and treatment issues. In G. L. Flett & P. L. Hewitt (Eds.),

*Perfectionism: Theory, research, and treatment* (pp. 5–31). Washington, DC: American Psychological Association.

Flett, G. L., Besser, A., Wang, J., Hewitt, P. L., Sherry, S. B., & Velyvis, V. (2003). *Perfectionism, dimensions of self-esteem, and depression: An analysis of self-liking and self-competence.* Manuscript submitted for publication.

Flett, G. L., Hewitt, P. L., Blankstein, K. R., & Mosher, S. W. (1995). Perfectionism, life events, and depressive symptoms: A test of a diathesis-stress model. *Current Psychology, 14*, 112–137.

Flett, G. L., Hewitt, P. L., Blankstein, K. R., & O'Brien, S. (1991). Perfectionism and learned resourcefulness in depression and self-esteem. *Personality and Individual Differences, 12*, 61–68.

Flett, G. L., Hewitt, P. L., Blankstein, K. R., Solnik, M., & Van Brunschot, M. (1996). Perfectionism, social problem-solving ability, and psychological distress. *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 14*, 245–275.

Flett, G. L., Hewitt, P. L., Oliver, J. M., & Macdonald, S. (2002). Perfectionism in children and their parents: A developmental analysis. In G. L. Flett & P. L. Hewitt (Eds.), *Perfectionism: Theory, research, and treatment* (pp. 89–132). Washington, DC: American Psychological Association Press.

Flett, G. L., Russo, F. A., & Hewitt, P. L. (1994). Dimensions of perfectionism and constructive thinking as a coping response. *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 12*, 163–179.

Flett, G. L., Sawatzky, D. L., & Hewitt, P. L. (1995). Dimensions of perfectionism and goal commitment: A further comparison of two perfectionism measures. *Journal of Psychopathology and Behavioral Assessment, 17*, 111-124.

Frost, R. O., & Henderson, K. J. (1991). Perfectionism and reactions to athletic competition. *Journal of Sport and Exercise Psychology*, 13, 323–335.

Frost, R. O., & Marten, P. A. (1990). Perfectionism and evaluative threat. *Cognitive Therapy and Research, 14*, 559-572.

Frost, R. O., Heimberg, R. G., Holt, C. S., Mattia, C. S., & Neubauer, A. L. (1993). A comparison of two measures of perfectionism. *Personality and Individual Differences, 14*, 119–126.

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14*, 449-468.

Frost, R. O., Trepanier, K. L., Brown, E. J., Heimberg, R. G., Juster, H. R., & Makris, G. S. (1997). Self monitoring mistakes among subjects high and low in perfectionistic concerns over mistakes. *Cognitive Therapy and Research, 21*, 209–222.

Frost, R. O., Turcotte, T. A., Heimberg, R. G., Mattia, J. I., Holt, C. S., & Hope, D. A. (1995). Reactions to mistakes among subjects high and low in perfectionistic concern over mistakes. *Cognitive Therapy and Research, 19*, 195–205.

Halmi, K. A., Sunday, S. R., Strober, M., Kaplan, A., Woodside, D. B., & Fichter, M., (2000). Perfectionism in anorexia nervosa: Variation by clinical subtype, obsessionality, and pathological eating behavior. *The American Journal of Psychiatry, 157*, 1799–1805.

Hamachek, D. E. (1978). Psychodynamics of normal and neurotic perfectionism. *Psychology, 15*, 27–33.

Hewitt, P. L., & Flett, G. L. (1991a). Dimensions of perfectionism in unipolar depression. *Journal of Abnormal Psychology, 100*, 98–101.

Hewitt, P. L., & Flett, G. L. (1991b). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of Personality and Social Psychology, 60*, 456–470.

Hewitt, P. L., & Flett, G. L. (1993). Dimensions of perfectionism, daily stress, and depression: A test of the specific vulnerability hypothesis. *Journal of Abnormal Psychology*, *102*, 58–65.

Hewitt, P. L., Flett, G. L., & Ediger, E. (1996). Perfectionism and depression: Longitudinal assessment of a specific vulnerability hypothesis. *Journal of Abnormal Psychology*, *105*, 276–280.

Hewitt, P. L., Flett, G. L., & Endler, N. S. (1995). Perfectionism, coping, and depression symptomatology in a clinical sample. *Clinical Psychology and Psychotherapy*, *2*, 47–58.

Hewitt, P. L., Flett, G. L., & Turnbull, W. (1992). Perfectionism and MMPI indices of personality disorder. *Journal of Psychopathology and Behavioral Assessment*, *14*, 323–335.

Hewitt, P. L., Flett, G. L., & Weber, C. (1994). Dimensions of perfectionism and suicide ideation. *Cognitive Therapy and Research*, *18*, 439–460.

Hewitt, P. L., Flett, G. L., Besser, A., Sherry, S. B., & McGee, B. (2003). Perfectionism is multidimensional: A reply to Shafran, Cooper, and Fairburn (2002). *Behaviour Research and Therapy, 41,* 1221-1236.

Hewitt, P. L., Flett, G. L., Norton, G. R., & Flynn, C. (1998). Dimensions of perfectionism and chronic symptoms of unipolar and bipolar depression. *Canadian Journal of Behavioural Science*, *30*, 234–242.

Hewitt, P. L., Flett, G. L., Turnbull-Donovan, W., & Mikail, S. (1991). The Multidimensional Perfectionism Scale: Reliability, validity, and psychometric properties in psychiatric samples. *Psychological Assessment*, *3*, 464–468.

Hewitt, P. L., Newton, J., Flett, G. L., & Callander, L. (1997). Perfectionism and suicide ideation in adolescent psychiatric patients. *Journal of Abnormal Child Psychology, 25*, 95–101.

Hill, R. W., McIntire, K., & Bacharach, V. R. (1997). Perfectionism and the big five factors. *Journal of Social Behaviour and Personality, 12*, 257–270.

Juster, H. R., Heimberg, R. G., Frost, R. O., Holt, C. S., Mattia, J. I., & Faccenda, K. (1996). Social phobia and perfectionism. *Personality and Individual Differences*, *21*, 403-410.

Kawamura, K. Y., Hunt, S. L., Frost, R. O., & DiBartolo, P. M. (2001). Perfectionism, anxiety, and depression: Are the relationships independent? *Cognitive Therapy and Research, 25*, 291–301.

Khawaja, N. G., & Armstrong, K. A. (2005) An investigation of the factor structure of the Frost Multidimensional scale on the basis of the Australian population. *Australian Journal of Psycholog, 57*, 129-138.

Lundh, L. G., Broman, J. E., Hetta, J., & Saboonchi, F. (1994). Perfectionism and insomnia. *Scandinavian Journal of Behaviour Therapy*, *23*, 3-18.

Lynd-Stevenson, R. M., & Hearne, C. M. (1999). Perfectionism and depressive affect: The pros and cons of being a perfectionist. *Personality and Individual Differences, 26*, 549-562.

Minarik, J. K., & Burns, L. R. (1998). Relations of eating and symptoms of depression and anxiety to the dimensions of perfectionism among undergraduate women. *Cognitive Therapy and Research, 20*, 155-169.

Minarik, M. L., & Ahrens, A. H. (1996). Relations of eating behavior and symptoms of depression and anxiety to the dimensions of perfectionism among undergraduate women. *Cognitive Therapy and Research, 20*, 155–169.

Mongrain, M. (1993). Dependency and self-criticism located within the five-factor model of personality. *Personality and Individual Differences, 15*, 455–462.

Powers, T. A., Koestner, R., & Topciu, R. A. (2005). Implementation intentions, perfectionism, and goal progress: Perhaps the road to hell is paved with good intentions. *Personality and Social Psychology Bulletin, 31*, 902–912.

Rhéaume, J., Freeston, M. H., Dugas, M. J., Letarte, H., & Ladouceur, R. (1995). Perfectionism, responsibility, and obsessive-compulsive symptoms. *Behaviour Research and Therapy*, *33*, 785-794.

Rice, K. G., & Dellwo, J. P. (2002). Perfectionism and self-development: Implications for college adjustment. *Journal of Counseling and Development, 80*, 190–196.

Rice, K. G., Ashby, J. S., & Slaney, R. B. (1998). Self-esteem as a mediator between perfectionism and depression: A structural equations analysis. *Journal of Counseling Psychology, 45,* 304–314.

Rogers, C. R. (1951). *Client-centered therapy: Its current practice, implications, and theory*. Boston: Houghton Mifflin.

Saboonchi, F., & Lundh, L. (2003). Perfectionism, anger, somatic health, and positive affect. *Personality and Individual Differences, 34,* 1-5.

Shafran, R., & Mansell, W. (2001). Perfectionism and psychopathology: A review of research and treatment. *Clinical Psychology Review, 21*(6), 879–906.

Shafran, R., Cooper, Z., & Fairburn, C. G. (2002). Clinical perfectionism: A cognitive behavioural models of the perfectionism construct. *Personality and Individual Differences, 36*, 1373–1385.

Srinivasagam, N. M., Kaye, W. H., Plotnicov, K. H., Greeno, C., Weltzin, T. E., & Rao, R. (1995). Persistent perfectionism, symmetry, and exactness after long-term recovery from anorexia nervosa. *The American Journal of Psychiatry, 152*, 1630–1634.

Steele, A., Corsini, N., & Wade, T. D. (2007). The interaction of perfectionism, perceived weight status, and self-esteem to predict bulimic symptoms: The role of 'benign' perfectionism. *Behaviour Research and Therapy, 45*, 1647–1655.

Stöber, J. (1998). The Frost Multidimensional Perfectionism Scale: More perfect with four (instead of six) dimensions. *Personality and Individual Differences, 24*(4), 481-491.

Terry-Short, L. A., Owens, R. G., Slade, P. D., & Dewey, M. E. (1995). Positive and negative perfectionism. *Personality and Individual Differences, 18*, 663-668.

Vredenburg, K., Flett, G. L., & Krames, L. (1993). Analogue versus clinical depression: A critical re-appraisal. *Psychological Bulletin*, *113*, 327–344.

Widiger, T. A., & Costa, P. T. Jr., (2002). Five-factor model personality disorder research. In P. T. Costa, Jr. & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 59–87). Washington, DC: American Psychological Association.

Zhang, Y., Gan, Y., & Cham, H. (2007). Perfectionism, academic burnout and engagement among Chinese college students: A structural equation modeling analysis. *Personality and Individual Differences, 43*, 1529–1540.

Zuroff, D. C. (1994). Depressive personality styles and the five-factor model of personality. *Journal of Personality Assessment, 63*, 453–472.

## About the Authors

**Joel C. Navarez** is a Registered Guidance Counselor and a Licensed Teacher. He finished his BS Psychology at Colegio de San Juan de Letran, Calamba and earned his MA Education major in Guidance and Counseling degree at Philippine Normal University. Currently, he is pursuing a Ph.D. in Clinical Counseling Psychology at De La Salle University – Manila.

**Ryan Francis O. Cayubit** is an Assistant Professor at the Department of Psychology of the University of Santo Tomas. He graduated from Colegio de San Juan de Letran, Manila with the degree BS Psychology. He earned his MA in Psychology from the UST Graduate School and is currently taking up his PhD in Educational Psychology (Quantitative Methods) at De La Salle University – Manila.

Educational Measurement and Evaluation Review (2011), Vol. 2, 92-110


http://pemea.club.officelive.com/ EMEReview.aspx

# Using Logistic Regression and Mantel-Haenszel Statistic in Differential Item Functioning Analysis: A Comparative Study

**Jose Q. Pedrajita**

*University of the Philippines, Diliman*
*jose.pedrajita@up.edu.ph*

**Abstract** This research article provides a demonstration of differential item functioning (DIF) analysis. DIF analysis investigates a differential characteristic of a test item between groups of examinees and is useful in detecting potentially biased items toward a particular group of examinees. The study made use of test scores of 200 junior high school students on a Chemistry Achievement Test, a measure tested for its psychometric properties. One hundred students came from a public school, while the other 100 were private school examinees; 100 students were males and the other 100 were females; and 95 students were of low ability and 105 students were of high ability based on their English II grades. Two non-parametric methods, the Logistic Regression and the Mantel-Haenszel Statistic, were applied in the DIF analysis to identify potentially biased test items between examinees matched on class type, gender, and English ability. Thereafter, the results for the two methods were compared. The findings revealed the presence of items indicating class type, gender, and English ability bias. There was a high degree of correspondence between the Logistic Regression and the Mantel-Haenszel Statistic in identifying potentially biased test items.

**Keywords:** *differential item functioning - differential item functioning analysis - item bias*

## Introduction

A critical step in the development of educational assessment instruments is to ensure that no individual or group responding to the instrument is disadvantaged in any way. This is an important process to achieve test equity. Test equity is primarily achieved by ensuring that a test measures only construct-relevant differences between subpopulation of examinees. If test equity is not achieved, a test or test item is biased toward a particular subpopulation of examinees (Kanjee, 2007).

*Bias* is not the mere presence of a score difference between groups. In test items, bias is the presence of a systematic error in measurement (Camilli & Shepard, 1994). Items may be judged to be more or less difficult for a particular group by comparing them with the performance of another group or groups drawn from the same population.

Test items are subjected to DIF detection techniques to determine whether or not they conform to a given set of psychometric rules in the same way for all persons in a population, regardless of any subgroup membership within that population.

One way to investigate bias at the item level is through *differential item functioning* (DIF) analysis. DIF analysis is a means of statistically identifying unexpected differences in performance across matched groups of examinees. It compares the performance of matched majority (or reference) and minority (or focal) group examinees.

*Differential item functioning* is said to be present in a test item when, despite controls for overall test performance, examinees from different groups have a different probability or likelihood of answering an item correctly or when examinees from two subpopulations with the same trait level have different expected scores on the same item (Camilli & Shepard, 1994; Kamata & Vaughn, 2004).Thus, an item that exhibits DIF may or may not be biased for or against any group (Kanjee, 2007). DIF may be attributed to item bias but may also reflect performance differences that the test is designed to measure (Camilli & Shepard, 1994).

According to O'Neill and McPeek (1993, p. 76 ), "The fundamental principle of DIF is simple: Examinees who know the same amount about a topic should perform equally well on an item testing that topic regardless of their sex, race, or ethnicity."

Methods for detecting DIF have proliferated in recent years. A researcher wishing to select a DIF method is confronted with many methods and with no clear guidelines for choosing among them. The of comparison DIF methods is an important practical concern, since both the size of sample required and the cost associated with the procedures differ widely. If all the DIF approaches were to identify the same items as biased, one could use the simplest and least expensive approach. However, if the approaches identify different items as biased, it becomes necessary to determine those methods which are most valid. Thus, there is a need to empirically compare the DIF methods. The present study represents an attempt to meet this need. It would test two non-parametric DIF methods, the Logistic Regression (LR) and the Mantel-Haenszel (MH) Statistic in detecting differential item functioning in a Chemistry Achievement Test.

Earlier studies which had used both LR and MH procedures to analyze data revealed that the two procedures yield very similar results with respect to uniform DIF (Mazor, Kanjee, & Clauser, 1995; Rogers & Swaminathan, 1993; Kamata, & Vaughn, 2004). These results also suggest that when identical matching criteria are used, the MH procedure and the LR produce extremely similar classifications. Thus, there is a high degree of correspondence between the LR and the MH procedures when either one or two ability estimates were included in the analysis. LR has shown that under comparable conditions, when matching is based on a single test score, it produces results that are extremely similar to those produced using the MH Statistic.

Findings in previous studies which have used both LR and the MH procedure revealed the following: *First*, the number of items exhibiting bias with both the LR and the MH procedures seem high. Apparently, both LR and MH are the most sensitive among the contingency table approaches. *Second*, consistent with earlier research, regardless of which criterion the comparison is based on, the MH and the LR procedures result in similar numbers of items (and similar items) being identified.

Studies which have used both LR and the MH procedure to analyze data have found that the two procedures yield very similar results with respect to uniform DIF. They reported substantial agreement between the two procedures.

In this study, these two methods for detecting DIF will be evaluated further in terms of external evidence of validity. The types of validity evidence for a DIF technique would be a demonstration that: a) the procedure is not selecting items at random; and (b) the results obtained with different methods tend to agree. Perfect agreement would probably not be expected, due to differences in the assumptions and limitations of the various methods.

Currently, there is a dearth of studies on comparison of DIF detection approaches in the literature on educational measurement, research, and evaluation in the Philippines.

## Methodology

This paper aimed to detect items indicating *class type*, *gender*, and *ability bias* in a Multiple Choice type Chemistry Achievement Test through differential item functioning analysis. The descriptive-comparative research design was used. The Chemistry Achievement Test was administered to three matched groups, namely: public and private, male and female, and low and high English ability examinees. Thereafter, the examinees' scores were subjected to the two DIF methods, the LR and MH Statistic, to identify items indicating class type, gender, and ability bias.

*Class type bias*, *gender bias*, and *ability bias* refers to the differing probabilities of success on an item(s) between the 100 public and the 100 private, the 100 male and the 100 female, and the 95 low and the 105 high ability examinees, respectively. For each matched group, the total number of examinees adds up to 200, which is the total sample in this study. These examinees were third year high school students taken from the top, middle, and lower class sections of a public and a private school in the Division of City Schools, Quezon City, Philippines.

The preparation of the Chemistry Achievement Test items involved the following steps: (1) development of a Table of Specifications; (2) consultation with experts; (3) generation of an item pool; (4) review of the initial item pool by experts; (5) field-testing; and (6) item analysis and test revision.

The following indices of item difficulty and item discrimination were used in deciding whether to discard or retain an item.

| ***Index of Difficulty*** | ***Index of Discrimination*** |
|---|---|
| 91% & above – very easy, to be discarded | .40 and up – very good item |
| 76 – 90 – easy, needs revision | .30 - .39 – good item |
| 26 – 75 – highly acceptable, optimum difficulty | .20 - .29 – marginal item |
| 11 – 25 – difficult, needs revision | .19 & below – poor item |
| 10% and below – very difficult, to be discarded | |

Items with difficulty indices within .20 to .80 and discrimination indices of .30 to .80 were retained. This means that items with difficulty level of .20 and below (very difficult) and .81 and above (very easy) were discarded. Likewise, items with discrimination indices of .20 to .29 (marginal item) and .19 and below (poor item) were rejected.

However, after the item analysis of the 75-item pool, 14 marginal items and 4 poor items were considered for inclusion in the final form to complete the required number of items of 50 in the research instrument. The basis of consideration is that their difficulty levels ranged from easy, optimum difficulty, to difficult which are acceptable difficulty ranges. The items considered for inclusion have discrimination indices of .20 to .29 (marginal items) and .19 and below (poor items). Their difficulty indices ranged from .20 to .80 which is the acceptable range of difficulty for test items.

Table 1
*Retained Items in the Chemistry Achievement Test*

| Former Item No. | Index of Difficulty | Index of Discrimination | New Item No. |
|---|---|---|---|
| 2 | .5 | .32 | 1 |
| 3 | .70 | .53 | 2 |
| 4 | .59 | .44 | 3 |
| 6 | .48 | .72 | 4 |
| 7 | .39 | .46 | 5 |
| 8 | .69 | .5 | 6 |
| 9 | .52 | .21 | 7 |
| 10 | .5 | .44 | 8 |
| 12 | .53 | .32 | 9 |
| 14 | .62 | .63 | 10 |
| 15 | .53 | .25 | 11 |
| 16 | .67 | .53 | 12 |
| 17 | .43 | .25 | 13 |
| 18* | .48 | .15 | 14 |
| 19 | .39 | .22 | 15 |
| 20 | .37 | .5 | 16 |
| 21 | .8 | .22 | 17 |
| 24 | .45 | .28 | 18 |
| 26 | .34 | .31 | 19 |
| 27 | .57 | .41 | 20 |
| 28 | .40 | .43 | 21 |
| 33 | .43 | .31 | 22 |
| 35 | .73 | .22 | 23 |
| 37 | .37 | .5 | 24 |
| 38 | .54 | .47 | 25 |
| 39 | .81 | .31 | 26 |
| 41 | .48 | .28 | 27 |
| 42 | .23 | .22 | 28 |
| 43 | .26 | .21 | 29 |
| 44 | .51 | .21 | 30 |
| 46 | .59 | .31 | 31 |
| 47 | .54 | .65 | 32 |
| 48 | .51 | .47 | 33 |
| 49* | .37 | .19 | 34 |
| 50 | .40 | .37 | 35 |
| 51 | .47 | .62 | 36 |
| 52 | .40 | .25 | 37 |
| 53 | .58 | .22 | 38 |
| 54 | .5 | .44 | 39 |
| 57 | .21 | .31 | 40 |
| 58 | .39 | .4 | 41 |
| 59 | .29 | .35 | 42 |
| 60 | .48 | .65 | 43 |
| 63* | .5 | .18 | 44 |
| 64 | .44 | .5 | 45 |
| 68 | .64 | .28 | 46 |
| 69 | .56 | .38 | 47 |
| 71 | .47 | .5 | 48 |
| 72* | .40 | .13 | 49 |
| 73 | .56 | .38 | 50 |

*Items with poor discrimination index but with highly acceptable difficulty index which were considered for inclusion in the final form of the Chemistry Achievement Test.

The poor items were items 18, 49, 63, and 72 with discrimination indices ranging from .19 and below, but their difficulty levels are all highly acceptable which is within the range of .26 to .75. The marginal items includes one easy item (item 21 with difficulty level of .8); one difficult item (item 42 with difficulty level of .23); and twelve

items of optimum difficulty (items 9, 15, 17, 19, 24, 35, 41, 43, 44, 52, 53, and 68) with difficulty indices ranging from .26.

Table 1 shows the retained items in the final form of the Chemistry Achievement Test.

Table 2 shows the discarded items in the Chemistry Achievement Test. These discarded items have difficulty indices below .20 (very difficult) and above .80 (very easy) and discrimination indices below .20.

Table 2
*Discarded Items in the Chemistry Achievement Test*

| Item No. | Index of Difficulty | Index of Discrimination |
|---|---|---|
| 1 | .89 | .16 |
| 5 | .17 | - .03 |
| 11 | .86 | .10 |
| 13 | .57 | .03 |
| 22 | .73 | - .09 |
| 23 | .18 | .37 |
| 25 | .7 | .22 |
| 29 | .19 | .06 |
| 30 | .14 | .22 |
| 31 | .35 | - .03 |
| 32 | .20 | - .03 |
| 34 | .37 | 0 |
| 36 | .51 | - .09 |
| 40 | .16 | 0 |
| 45 | .29 | .09 |
| 55 | .26 | .15 |
| 56 | .39 | - .16 |
| 61 | .29 | - .15 |
| 62 | .45 | - .03 |
| 65 | .44 | - .44 |
| 66 | .12 | 0 |
| 67 | .14 | .16 |
| 70 | .23 | .03 |
| 74 | .20 | .03 |
| 75 | .42 | - .03 |

Table 3 shows the content area,; skills measured, and the number, percentages and placement of items in the Chemistry Achievement Test.

The instrument used in this study, the Chemistry Achievement Test, was composed of 50 items. These items were taken from five instructional units and were classified according to different levels of the cognitive domain: 5 or 10 % were *knowledge level* questions; 4 or 8 % were *comprehension level* questions; 13 or 26 % were *application level* questions; 17 or 34 % were *analysis level* questions; 7 or 14 % were *synthesis level* questions; and 4 or 8 % were *evaluation level* questions. The cognitive levels ranged from simple to complex. These questions were taken from the different learning competencies in Chemistry for a whole school year.

Unit 1 deals with introductory concepts in Chemistry composed of three chapters. Unit 2 is about the concept of matter consisting of three chapters dealing with *behavior of molecules*, *a view of the atom*, and *atoms in the periodic table*. Unit 3 deals with why and how atoms combine. It is composed of two chapters dealing with *bond formation* and *shape of molecules*. Unit 4 deals with the factors affecting chemical reactions. It has three chapters, two of which deal with *chemical activities* and the other with *chemical equilibrium*. Unit 5 deals with how chemistry creates new technologies. It

is composed of four chapters dealing with *solutions; acids, bases, and salts*; *colloids*; and *life and carbon compounds.*

Table 3
*Item Content of the Chemistry Achievement Test*

| Cognitive Domain | C O N T E N T Unit I | Unit II | Unit III | Unit IV | Unit V | Total Items | % | Item Placement |
|---|---|---|---|---|---|---|---|---|
| Knowledge | | 1, 2 | 3 | 4 | 5 | 5 | 10 | 1 - 5 |
| Comprehension | 6 | 7 | 8 | | 9 | 4 | 8 | 6 - 9 |
| Application | 10, 11 | 12, 13, 14 15 | 16, 17, 18 | 19, 20, 21 | 22 | 13 | 26 | 10 - 22 |
| Analysis | 23,24,25 26 | 27,28,29 30 | | 31,32,33 34, 35 | 36,37,38 39 | 17 | 34 | 23 - 39 |
| Synthesis | 40,41 | 42,43 | | 44,45 | 46 | 7 | 14 | 40 - 46 |
| Evaluation | 47 | 48,49 | 50 | | | 4 | 8 | 47 - 50 |
| No. of Items | 10 | 15 | 6 | 11 | 8 | 50 | 100% | |
| Percent | 20% | 30% | 12% | 22% | 16% | 100% | | |

Of the 5 knowledge level items, none was taken from Unit 1; items 1 and 2 were taken from Unit 2; item 3 was taken from Unit 3; item 4 was taken from Unit 4; and item 5 was taken from Unit 5.

Of the 4 comprehension items, item 6 was taken from Unit 1; item 7 was taken from Unit 2; item 8 was taken from Unit 3; no item was taken from Unit 4; and item 9 was taken from Unit 5.

Of the 13 application level items, items 10 and 11 were taken from Unit 1; items 12, 13, 14, and 15 were taken from Unit 2; items 16, 17, and 18 were taken from Unit 3; items 19, 20, and 21 were taken from Unit 4; and item 22 was taken from Unit 5.

Of the 17 analysis level items, items 23, 24, 25, and 26 were taken from Unit 1; items 27, 28, 29, and 30 were taken from Unit 2; no item was taken from Unit 3; items 31, 32, 33, 34, and 35 were taken from Unit 4; and items 36, 37, 38, and 39 were taken from Unit 5.

Of the 7 synthesis level items, items 40 and 41 were taken from Unit 1; items 42, and 43 were taken from Unit 2; no item qualifies in Unit 3; items 44 and 45 were taken from Unit 4; and item 46 was taken from Unit 5.

Of the 4 evaluation items, item 47 was taken from Unit 1; items 48 and 49 were taken from Unit 2; and item 50 was taken from Unit 3. No item from Unit 4 and 5 qualified.

As per instructional unit, 10 or 20 percent of the items belong to Unit 1; 15 or 30 percent of the items were taken from Unit 2; 6 or 12 percent of the items came from Unit 3; 11 or 22 percent of the items belong to Unit 4; and 8 or 16 percent of the items belong to Unit 5.

Unit 1 is composed of one comprehension item (item 6); two application items (items 10 and 11); four analysis items (items 23, 24, 25, and 26); two synthesis items (items 40 and 41); and one evaluation item (item 47).

Unit 2 is composed of two knowledge level items (items 1 and 2); one comprehension item (item 7); four application items (items 12, 13, 14, and 15); four analysis items (items 27, 28, 29, and 30); two synthesis items (items 42 and 43); and two evaluation items (items 48 and 49).

Unit 3 is composed of one knowledge level item (item 3); one comprehension item (item 8); three application items (items 16, 17, and 18); no analysis and synthesis items; and one evaluation item (item 50).

Unit 4 is composed of one knowledge level item (item 4); no comprehension item; three application items (items 19, 20, and 21); five analysis items (items 31, 32, 33, 34, and 35); two synthesis items (items 44 and 45); and no evaluation item.

Unit 5 is composed of one knowledge item (item 5); one comprehension item (item 9); one application item (item 22); four analysis items (items 36, 37, 38, and 39); one synthesis item (item 46); and no evaluation item.

The data gathering procedures involved: (1) administration of the test to the public and private school examinees; and (2) checking and scoring the test. Whereas, the data analysis procedure includes: (1) assigning and matching of test papers to the three comparison groups by section and total score; (2) organizing data for every item into a three-way contingency table; (3) encoding data in the Statistical Analysis System (SAS) computer program; (4) LR and MH analyses for detecting and testing for differential item functioning/for each comparison group; and (5) identifying potentially biased test items; (6) comparing the results for the two methods.

In the LR analysis the predictor variables are the (a) *score interval* and *class type* for the public and private school examinees; (b) *score interval* and *sex* for the male and female examinees; and (c) *score interval* and *English ability* for the low and high ability examinees. The dependent variable for each matched group of examinees is *the odds of getting the item right*. A significant *score interval* indicates that examinees with a higher total score tend to score better in the examination. Likewise, a significant *class type, sex, and ability* indicates that the odds of getting an item right are different between the matched groups. The null hypothesis is that *for two groups at level j, the population value is zero for either the difference between the proportions correct or the log odds ratio*.

The MH analysis yields a chi-square test with one degree of freedom to test the null hypothesis that *there is no significant relationship between group membership and test performance over all items after controlling for total test score between the matched groups of examinees*.

To achieve statistical significance in both analyses, the computed chi-square value must be greater than the critical chi-square value of 3.84 and its probability should be less than the set alpha level of 0.05.

## Differential Item Functioning Analysis

**Class Type Bias.** Table 4 shows the items indicating bias between the public and the private school examinees identified in the LR and MH analyses.

Table 4
*Potentially Biased Items Identified in the Public/Private School Matched Examinees*

| Items | Concept/Skills Measured | LRX² | MHX² | Biased Against |
|---|---|---|---|---|
| 1 | gas property illustrated by garbage smoke entering the house | 11.65* | 11.48* | Private |
| 2 | element with Latin name "aurum" | 4.56* | 4.42* | Public |
| 3 | chemical bonds which held together two atoms in a molecule by the transfer of an electron from one atom to the other | 8.76* | 8.56* | Private |
| 5 | Filipino scientist who pioneered in the use of biogas/biomass as a source of energy | 49.41* | NS | Private |
| 8 | definition of valence electrons | 12.94* | 12.36* | Public |
| 9 | description of dialysis | 19.61* | 19.20* | Private |
| 10 | volume of a cube | 3.92* | 3.84* | Private |
| 13 | new pressure of the gas when the volume is compressed to a smaller quantity | 14.05* | 13.56* | Public |
| 14 | problem-solving on Boyle's Law | 5.33* | 5.15* | Public |
| 16 | how the chemical and molecular formula of sodium sulfate is correctly written | 6.41* | 6.24* | Public |
| 19 | solving for the molar mass of $Fe_2O_3$ | 5.34* | 5.30* | Private |
| 21 | the mass of oxygen in sulfur trioxide if the ratio of sulfur to oxygen is 2 : 3 with sulfur having a mass of 6 grams | 8.97* | 8.79* | Private |
| 22 | volume conversion | 24.84* | 23.39* | Public |
| 26 | indicators of chemical change | 5.08* | 4.98* | Private |
| 30 | correct position of Chlorine in the periodic table | 11.05* | 10.89* | Private |
| 32 | which of the given chemical equations is balanced | 7.58* | 7.26* | Public |
| 33 | identify the reactants in the given chemical equation | 7.07* | 6.94* | Private |
| 36 | classification of a solution which changes red litmus paper to blue | 4.97* | 4.83* | Public |
| 37 | factors which increases the solubility of a solute | 6.15* | 5.92* | Public |

Cont. Table 4

| | | | | |
|---|---|---|---|---|
| 40 | evidences of chemical change | 5.90* | 5.71* | Public |
| 41 | laws which govern changes in matter | 4.57* | 4.43* | Public |
| 46 | which two are components of a solution | NS | 3.51* | Public |
| 47 | strategy which is most probable in proving the given hypothesis in the given experiment | 16.08* | 15.76* | Private |

*p < .05 NS – not significant

The DIF analyses between the public and the private matched group identified 23 potentially biased items, 11of which indicates bias against the private school examinees, while 12 indicate bias against the public school examinees.

**Potentially Biased Items Against the Private School Examinees**. The 11 items indicating bias against the private school examinees were items 1, 3, 5, 9, 10, 19, 21, 26, 30, 33, and 47. In each of these items, the odds of getting the item right favored the public school examinees. That is, the private school examinees were less likely to be familiar with the concepts reflected in these items.

Item 1, which inquires as to what property of gases was illustrated by a burning garbage smoke entering the house, has a difficulty index of .59 (optimum) and discrimination index of .58 (very good). In this item, the public school examinees got 63 correct, whereas the private school examinees got 34 correct responses.

Item 3 asks to identify the chemical bonds that held together two atoms in a molecule by the transfer of an electron from one atom to the other. This item has a difficulty index of .85 (easy) and discrimination index of .24 (marginal). In this item, the public school examinees got 74 correct, while the private school examinees got 51 correct responses.

Item 5 asks to identify the Filipino scientist who pioneered the use of biogas/biomass as a source of energy. It has a difficulty index of .76 (easy) and discrimination index of .12 (poor). In this item, the public school examinees got 78 correct, while the private school examinees got 27 correct responses.

Item 9 asks to identify the correct description of dialysis from the given alternatives. It has a difficulty index of .73 (optimum) and discrimination index of .18 (poor). In this item, the public school examinees got 79 correct, whereas the private school examinees got 45 correct responses.

Item 10 is a problem-solving item about computing the volume of a given cube. It has a difficulty index of .7 (optimum) and discrimination index of .49 (very good). The public school examinees obtained 72 correct responses, while the private school examinees obtained 54 .

Item 19 inquires as to which of the given options is the molar mass of $Fe_2O_3$ (Fe = 56, O = 16). It has difficulty index of .5 (optimum) and discrimination index of 0.4 (very good). It is also a problem-solving item. In this item, the public school examinees got 44 correct responses, while the private school examinees got 24.

Item 21, which asks to identify from the given alternatives the mass of oxygen in sulfur trioxide if the ratio of sulfur to oxygen is 2:3 with sulfur having a mass of 6 grams, has a difficulty index of .78 (easy) and discrimination index of .27 (marginal). In this

item, the public school examinees got 75 correct responses, while the private school examinees got 51.

Item 26 inquires as to which of the given alternatives indicates a chemical change. Its difficulty index is .84 (easy) while its discrimination index is .21 (marginal). The public school examinees got 84 correct responses, whereas the private school examinees got 67.

Item 30 inquires as to which among given alternatives is the correct position of chlorine in the periodic table. Its difficulty index is .58 (optimum) and its discrimination index is .55 (very good). In this item, the public school examinees got 59 correct responses, while the private school examinees got 31 correct responses.

Item 33 asks to identify from among the given alternatives the reactants in the equation $4Fe + 3O_2 \longrightarrow 2Fe_2O_3$. Its difficulty index is .74 (optimum) and its discrimination index is .39 (good). In this item, the public school examinees got 79 correct responses, while the private school examinees got 58.

Item 47 asks to identify from the given alternatives the strategy which is most probable in proving the given hypothesis in the experiment. It has a difficulty index of .78 (easy) and discrimination index of .27 (marginal). The public school examinees got 79 correct responses, while the private school examinees got 48.

**Potentially Biased Items Against the Public School Examinees.** The 12 items indicating bias against the public school examinees were items 2, 8, 13, 14, 16, 22, 32, 36, 37, 40, 41 and 46. In each of these items, the odds of getting the item right favored the private school examinees. That is, the information reflected in these items was less likely to be familiar to the public school examinees.

Item 2, which inquires as to which of the given chemical elements has the Latin name "aurum," has a difficulty index of .79 (easy) and discrimination index of .36 (good). In this item, the private school examinees obtained 88 correct responses, while, the public school examinees got 79.

Item 8 asks to identify as to which of the given descriptions refers to valence electrons. It has a difficulty index of .54 (optimum) and discrimination index of - .19 (poor). In this item, the private school examinees obtained 53 correct responses, while the public school examinees obtained 32.

Item 13 is a problem-solving item. It asksg as to which among given alternatives is the new pressure of the given gas when the volume is compressed to a smaller quantity. It has a difficulty index of .27 (optimum) and discrimination index of .18 (poor). The public school examinees got 21 correct responses, while the private school examinees got 45. To answer this item correctly, examinees need familiarity with the concept of volume-pressure relationship.

Item 14 is a problem-solving item about Boyle's Law. It has a difficulty index of .46 (optimum) and discrimination index of .49 (very good). In this item, the private school examinees obtained 57 correct responses, while the public school examinees got 45.

Item 16 asks to identify the correct chemical and molecular formula of sodium sulfate from the given alternatives. It has a difficulty index of .16 (difficult) and discrimination index of -.09 (negative). In this item, the public school examinees got 17 correct responses, while the private school examinees got 33.

Item 22 is a problem-solving item about volume calculation. It has a difficulty index of .36 (optimum) and discrimination index of .43 (very good). The public school examinees got 29 correct responses, while the private school examinees got 59.

Item 32 inquires as to which of the given chemical equations is balanced. Its difficulty index is .34 (optimum) while its discrimination index is .43 (very good). The private school examinees obtained 45 correc responsest, while the public school examinees got 34 .

Item 36 asks which option is the classification of a solution which changes red litmus paper to blue. It has a difficulty index of .39 (optimum) and discrimination index of 0 (zero). In this item, the public school examinees got 35 correct responses, while the private school examinees got 50.

Item 37 asks to identify which of the given factors increases the solubility of a solute. It has a difficulty index of .38 (optimum) and discrimination index of .4 (very good). In this item, the public school examinees got 37 correct responses, while the private school examinees got 49.

Item 40 asks to pick out from the given properties the evidences of chemical change. Its difficulty index is .21 (difficult) and its discrimination index is .3 (good). In this item, the private school examinees got 42 correct responses, while, the public school examinees got 28.

Item 41 inquires which of the three given gas laws governs changes in matter. Its difficulty index is .28 (optimum) while its discrimination index is .21 (marginal). In this item, the public school examinees got only 32 correct responses, while the private school examinees got 43.

Item 46 asks which of the two given options are the components of a solution. Its difficulty index is .73 while its discrimination index is .42. In this item, the public school examinees obtained 73 correct responses, while the private school examinees obtained 81.

A clear pattern was revealed in the analyses that items indicating bias against the private school examinees were relatively easier items, mostly belonging to the middle and the upper ranges of difficulty level while items indicating bias against the public school examinees were relatively difficult items, mostly belonging to the middle and lower ranges of difficulty levels.

It could be gleaned that the LR and the MH Statistic yielded very similar results. Both identified 22 biased items, 21 of which were identical items, except for item 5 for the LR and item 46 for the MH chi square.

## Gender Bias

Table 5 shows the items indicating bias between the male and the female examinees which were identically identified in the LR and MH analyses.

Both statistical analyses identified seven identical items indicating bias between the male/female matched groups. They were items 1, 3, 17, 27, 34, 42, and 47.

Table 5
*Potentially Biased Items Identified in the Male/Female Matched Examinees*

| Items | Concept/Skills Measured | LRX² | MHX² | Biased Against |
|---|---|---|---|---|
| 1 | gas property illustrated by garbage smoke entering the house | 4.92* | 4.79* | Male |
| 3 | chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other | 4.59* | 4.48* | Male |

Cont. Table 5

| | | | | |
|---|---|---|---|---|
| 17 | electron configuration of the element Sodium | 6.58* | 6.33* | Female |
| 27 | options which illustrates the compressibility of gases | 9.57* | 9.25* | Female |
| 34 | definition of reaction reversibility | 8.54* | 8.28* | Female |
| 42 | principles of Kinetic Molecular Theory | 5.63* | 5.48* | Male |
| 47 | strategy which is most probable in proving the given hypothesis in the given experiment | 5.26* | 5.12* | Male |

*$p < .05$

**Potentially Biased Items Against the Male Examinees.** Four items (1, 3, 42, and 47) indicates bias against the male examinees. In each of these items, the odds of getting the item right favored the female examinees. That is, the concept reflected in each of these items was less likely to be familiar to the male examinees.

Item 1, formerly identified as indicating bias against the private school examinees in both the LR and MH analyses, was also identified as indicating bias against the male examinees in the same analyses. The male examinees obtained 41 correct responses, while the female examinees obtained 56.

Item 3, which asks for the type of chemical bond which links two atoms in a molecule held together by the transfer of an electron from one atom to the other, was identified as indicating bias against the private school examinees both in the LR and MH analyses. Likewise, it was also identified as indicating bias against the male examinees in the same analyses. In this item, the male examinees obtained 55 correct responses, while the female examinees obtained 70 correct.

Item 42 asks to choose from the given concepts those which refer to the Kinetic Molecular Theory. Its difficulty index is .42 (optimum) while its discrimination index is .25 (marginal). In this item, the male examinees got 34 correct responses, while the female examinees got 50. The item content may have involved ideas that are more familiar to the female examinees.

Item 47, formerly identified as indicating bias against the private school examinees in both the LR and MH analyses, was also identified as indicating bias against the male examinees in the same analyses. It has a difficulty index of .78 (easy) and discrimination index of .27 (marginal). The female examinees obtained 71 correct responses, while the male examinees obtained 56.

**Potentially Biased Items against the Female Examinees.** Three items, 17, 27, and 34, were found to indicate bias against the female examinees. In each of these items the odds of getting the item right favored the male examinees. That is, the concept reflected in each of these items was less likely to be familiar to the female examinees.

Item 17 asks to choose the correct electronic configuration of the element $_{47}$Na from the given alternatives. It has a difficulty index of .86 (easy) and discrimination index of .27 (marginal). In this item, the male examinees got 93 correct as against 81 correct responses of the female examinees.

Item 27 inquires as to which of the given statements illustrate the compressibility of gases. Its difficulty index is .58 (optimum) and its discrimination index is .19 (poor). In this item, the male examinees got 50 correct, while the female examinees got 30.

Item 34 asked to identify which of the given statements is true of reaction reversibility. Its difficulty index is .18 (difficult) while its discrimination index is 0 (poor). In this item, the males obtained 44 correct responses, while the females got 25.

The result of the analyses reveals that the LR and the MH Statistic were very similar. The LR and the MH Statistic have each identified identical and similar number of items.

In the study "Using Logistic Regression and the Mantel-Haenszel with Multiple Ability Estimates to Detect Differential Item Functioning," Mazor, Kanjee, and Clauser (1995) examined the effect of incorporating more than one ability estimate into a DIF analysis. Two different achievement tests were analyzed using LR, conditioning first on total score only, and second with both total score and an external ability estimate included in the equation. To provide a basis for comparison, MH analyses were also conducted. The specific hypothesis tested in this study was that adding a second (in this case an external) ability estimate would more completely define the latent ability space, and thus reduce the number of items showing significant group differences. The .01 level of significance was used in all analyses. It was anticipated that inclusion of additional ability estimates would have less of an effect on the latter comparison, as there would be little, if any, difference between males and females in verbal ability. Analyses of the M/F data sets for both tests revealed no substantial reduction in the number of items exhibiting DIF when the SAT-V score or the SAT-M score was included in the LR equation. This was true for both the Chemistry and the History tests. For the History test, the number of items exhibiting DIF increased by 2. When SAT-V or SAT-M was substituted for achievement test score, as the single conditioning variable, the number of items exhibiting significant DIF increased across both groups and both tests. The MH results showed a pattern remarkably similar to that for the LR analysis. The results provide evidence that including the second irrelevant ability estimate in the equation does not improve "matching" and therefore does not reduce the number of items flagged by DIF.

In the study "Improving the Matching for DIF Analysis by Conditioning on Both Test Score and an Educational Background Variable," Clauser, Nungester, and Swaminathan (1996) examined the potential to improve matching by conditioning simultaneously on test score and a categorical variable representing the educational background of the examinees. The data set comprises the responses of males and females to 440 items from a test of physicians' clinical skills. The background variable represented the area of the physician's residency training (i.e. internal medicine, surgery, etc.). The analysis was carried out using logistic regression and the Mantel-Haenszel statistic. These procedures were chosen because previous research has shown that under comparable conditions (when matching is based on a single test score) they produce results that are extremely similar. The hypothesis examined was that adding a conditioning variable would reduce the number of items flagged by DIF. The presumed mechanism of this reduction was that adding the variable would partially account for those differences in group performance on secondary abilities that were related to educational background. The initial analysis, in which examinees were matched on total test score, identified 132 items as displaying DIF. This represents approximately 30% of the items studied. When the additional variable was included in the model so that matching was based on both total test score and the educational background variable, the number of items identified was reduced to 84 (approximately 19%), a reduction of

more than one third in the number of items identified. When the analysis was repeated using the sub-scores instead of the total test score, similar results were obtained. The initial analysis identified 127 items (approximately 29%). When the background conditioning variable was added, 82 items were identified (approximately 19%). Again this represents a reduction of approximately one third in the number of items identified. The main finding supports the hypothesis that the addition of a variable representing educational background may improve matching and lead to a reduction in the number of items identified as displaying DIF.

## Ability Bias

Table 6 shows the items indicating bias between the low and the high ability examinees which were identically identified in the LR and MH analyses.

Table 6
*Potentially Biased Items Identified in the Low/High Ability Matched Examinees*

| Items | Concept/Skills Measured | $LRX^2$ | $MHX^2$ | Biased Against |
|---|---|---|---|---|
| 2 | element with Latin name "aurum" | 10.33* | 10.04* | Low |
| 3 | chemical bond which held together two atoms in a molecule by the transfer of an electron from one atom to the other | 14.63* | 14.65* | Low |
| 6 | scope of chemistry | 19.73* | 19.06* | Low |
| 8 | definition of valence electrons | 9.41* | 9.23* | Low |
| 13 | new pressure of the gas when the volume is compressed to a smaller quantity | 6.14* | 5.97* | Low |
| 19 | solving for the molar mass of $Fe_2 O_2$ | 17.90* | 17.47* | Low |
| 22 | volume conversion | 4.19* | 4.15* | Low |
| 29 | valence electrons of the Chlorine atoms | 4.98* | 4.81* | High |
| 36 | classification of a solution which changes red litmus paper to blue | 4.15* | 4.05* | Low |
| 38 | in which solution water is a solute | 6.73* | 6.45* | High |
| 45 | in which situation the process of oxidation is common | 4.43* | 4.22* | High |
| 48 | correct formula in solving for the new volume of the gas | 7.13* | 7.21* | Low |
| 50 | factor which causes the nails to rust | 7.51* | 7.57* | Low |

*$p < .05$

Both the LR and MH analyses identically identified 13 potentially biased itemsten of which( items 2, 3, 6, 8, 13, 19, 22, 36, 48, and 50) indicate bias against the low ability examinees, while only three, items 29, 38, and 45, indicate bias against the high ability examinees.

**Potentially Biased Items against the High Ability Examinees**. Three items, 29, 38, and 45, were identically identified as indicating bias against the high ability examinees in both the LR and MH analyses. In each of these items, the odds of getting the item right favored the low ability examinees. That is, the high ability examinees were less likely to be acquainted with the concept reflected in these items.

Item 29 inquires as to how many valence electrons does a Chlorine atom have. Its difficulty index is .38 (optimum) while its discrimination index is .28 (marginal). In this item, the low ability examinees obtained 43 correct responses, while the high ability examinees got 39. The low ability examinees outscored the high ability examinees.

Item 38 asks in which of the given solutions is water a solute. Its difficulty index is .41 (optimum) while the discrimination index is .08 (poor). The low ability examinees obtained 42 correct responses, while the high ability examinees got 37.

Item 45 asks in which of the four given phenomena is the process of oxidation is common. It has a difficulty index of .5 (optimum) and discrimination index of .34 (good). In this item, the high ability examinees obtained 52 correct responses, while the low ability examinees got 48.

**Potentially Biased Items Against the Low Ability Examinees**. Ten items, 2, 3, 6, 8, 13, 19, 22, 36, 48, and 50 favored the high ability examinees. That is, the concept reflected in these items was less likely to be familiar to the low ability examinees.

Item 2, formerly identified as indicating bias against the public school examinees in both LR and MH analyses, was also identified as indicating bias against the low ability examinees in the same analyses. In this item, the low ability examinees got 71 correct responses, while the high ability examinees got 96.

Item 3, previously identified as potentially biased against the private and the male examinees in both the LR and MH analyses, was likewise identified by both methods as potentially biased against the low ability examinees. In this item, the low ability examinees got 43 correct responses, while the high ability examinees got 82.

Item 6 inquires as to which of the given concepts is within the scope of chemistry. It has a difficulty index of .79 (easy) and discrimination index of .42 (very good). In this item, the low ability examinees got 59 correct responses as against 93 of the high ability examinees.

Item 8, potentially biased against the public school examinees in both analyses, is again identified as indicating bias against the low ability examinees in the same analyses. This item is asking to identify the definition of valence electrons among the given options. In this item, the low ability examinees got 27 correct responses, while the high ability examinees got 58.

Item 13, previously identified as potentially biased against the public in the LR and MH analyses, was also identified as potentially biased against the low ability examines in the same analyses. In this item, the low ability examinees got 24 correct responses, while the high ability examinees got 42.

Item 19, indicating bias against the private school examinees, was likewise identified as indicating bias against the low ability examinees in both analyses. This item asked for the molar mass of the given compound from the given options. In this item, the low ability examinees got 15 correct responses, while the high ability examinees got 53. The low ability examinees scored miserably on the item.

Item 22, identified as potentially biased against the public school examinees, was also identified as indicating bias against the low ability examinees in the same analyses. The scores in this item were 31 correct for the low ability as against 56 correct for the high ability examinees.

Item 36, indicating bias against the public school examinees, was also identified as indicating bias against the low ability examinees in the same analyses. This item asked for the classification of a solution that changes red litmus paper to blue. The high ability examinees obtained 53 correct responses, while the low ability examinees got 33.

Item 48 asks to choose the correct formula in solving for the new volume of the gas in the given problem. Its difficulty index is .7 (optimum) while its discrimination index is .55 (very good). The scores in this item were 36 correct responses for the low ability as against 74 correct responses of the high ability examinees. This item was heavily indicating bias against the low ability examinees.

Item 50 asks to identify which of the given factors causes nails to rusts. It has a difficulty index of .66 (optimum) and discrimination index of .61 (very good). The scores in this item were 44 correct for the low ability examinees as against 82 for the high ability examinees.

The LR and the MH Statistic yielded very similar results. Each identified 13 identical potentially biased items.

A closer scrutiny of the items indicating bias against the low ability examinees shows that the these items have difficulty indices ranging from optimum difficulty to very easy. On the other hand, the items indicating bias against the high ability examinees have difficulty indices only within the optimum difficulty level. Hence, the pattern of bias was heavily against the low ability examinees.

The test was relatively difficult for the low ability examinees. The two groups did not have equal opportunities in learning experience related to the potentially biased items. The wording of the question may also be unfamiliar to the disadvantaged group. The performance differences in the low/high matched groups may have been due to the differences in ability to interpret or comprehend written English. Language ability is relevant to the purpose of testing. It is an important dimension on any test which requires more than the most minimal reading.

Thus, in the LR analysis, the null hypothesis that *the population value is zero for either the difference between the proportions correct or the log odds ratio on the test items between the public and private, the male and female, and the low and high ability examinees* is rejected. In the MH analysis, the null hypothesis that *there is no significant relationship between group membership and test performance on the test items between the public and private, the male and female, and the low and high ability examinees* is likewise rejected in favor of the alternative hypothesis.

The number of items exhibiting bias with both LR and MH procedures seem high. Apparently, both LR and MH are very sensitive DIF methods. Consistent with earlier research, regardless of which criterion the comparison is based on, the two procedures result in similar numbers of items (and identical items) being identified (Rogers & Swaminathan, 1993).

The findings in this study confirmed those of earlier studies which had used both LR and MH procedures to analyze data that the two procedures yield very similar results with respect to DIF. These results also suggest that when identical matching criteria are used, the MH procedure and the LR produce extremely similar classifications. Thus, there is a high degree of correspondence between the LR and the MH procedures when either one or two ability estimates were included in the analysis. LR has shown that under comparable conditions, when matching is based on a single test score, it produces results that are extremely similar to those produced using the MH Statistic.

The two methods for detecting DIF demonstrated that: a) the procedure does not select items at random; and (b) the results obtained with the two methods tend to

agree. Perfect agreement, however, was not expected, due to differences in the assumptions and limitations of the two methods.

The presence of class type, gender, and ability bias can be attributed to the: 1) discrepancies in the curriculum of the public and private school; 2) unfamiliarity with the content of the items which caused examinees to be attracted to the incorrect options; 3) ambiguities in the item stem, keyed response, or distracter; 4) disparities in the matched examinees' exposure to vocabularies, concepts or skills reflected on the items; 5) inability of the matched examinees to comprehend or understand the concepts reflected on the items.

Mazor, Kanjie, and Clauser (1995) examined the effect of incorporating more than one ability estimate into a DIF analysis. Two different achievement tests, Chemistry and History, were analyzed using LR, conditioning first on total score only, and second with both total score and an external ability estimate included in the equation. To provide a basis for comparison, MH analysis was also conducted. The specific hypothesis tested in the study was that adding a second ability estimate would more completely define the latent ability space, and thus reduce the number of items showing significant differences. In the comparison between examinees who reported English as their best language (EBL) and those who reported some other language as their best language (OBL), the results revealed that when a second ability estimate (SAT-V) was added into the LR analysis for the EBL/OBL comparison, the number of items exhibiting DIF decreased substantially as compared to the number exhibiting DIF when only the total score was used as the criterion. On both tests, when the SAT-M score was used instead of the SAT-V score, a slight increase in the number of items exhibiting DIF was observed as compared to using only the total score. The MH results showed a pattern remarkably similar to that in the LR analysis. The pattern of results was consistent with the hypothesis that at least some of the performance differences in the reference/focal group comparison were due to differences in ability in interpreting or comprehending written English.

## Conclusions

The results of the differential item functioning analysis showed that there were statistically biased test items between the public and the private, the male and the female, and the low and the high ability examinees. Hence, *class type*, *gender*, and *ability bias* are present in the Chemistry Achievement Test. Overall, it appears that the public school examinees performed better than the private school examinees; male and female examinees performed fairly; and low ability examinees performed more miserably than the high English ability examinees. Ability bias was heavily tilted against the low ability examinees.

Investigating bias at the item level is particularly useful in the process of test development, in which biased items are revised or removed. This is a legitimate and important process in an attempt to achieve test equity (Kamata & Vaughn, 2004). Test equity is primarily achieved by ensuring that a test measures only construct-relevant differences between subpopulations (Messick, 1989 as cited in Kamata & Vaughn, 2004). If test equity is not achieved, a test or test item is biased toward a particular subpopulation of the test taking population. Statistically, a test or test item is said to be biased if the expected test or item scores are not the same for examinees from different subpopulations, given the same level of trait that the test intends to measure (Kamata & Vaughn, 2004).

In deciding which DIF method to use, it is appropriate to choose methods which are most valid. Valid methods may be very sensitive and may have a very high detection rate in identifying biased test items. But it is better for test development for it could identify all items which are possibly biased, and then eliminate or revise such biased items in order to purify and maintain the measurement qualities of the test.

On the other hand, if methods which may not be so sensitive and with a very low detection rate are used, some items which could be possibly biased may not be identified and may remain part of the test content, thereby, still affecting and contaminating the validity and reliability of the test.

`Item bias methods with high detection rate are preferable over those with low detection rate in purifying assessment instrument. That is, test items should be free of bias.

Findings can significantly contribute to educational measurement. The use of statistical methods in identifying biased test items is a relatively better kind of item analysis. By subjecting test items to item bias detection approaches, test items which were unfairly difficult and widely discriminating for a particular group of examinees are determined. By eliminating, replacing, or revising these biased items a valid, reliable, and fairer test can be made.

## Recommendations

Test experts and developers should use contingency table (CT) methods, particularly Logistic Regression and Mantel-Haenszel Statistic, in item bias detection. These two methods are viable in the detection of DIF and are widely implemented in both test construction and research settings.

Educational evaluation practitioners engaged in the development of assessment tools can use Logistic Regression or Mantel-Haenszel Statistic for bias correction. That is, biased items should be revised or replaced to refine and purify the required item content of a test. This process could make differentially functioning items between groups of interest be more valid, reliable, and fair. Bias correction can improve the validity and reliability of a test.

In this study, matching was done by conditioning simultaneously on test score, and a categorical variable, namely, *class type* for the public/private comparison group, *sex* for the male/female comparison group, and *English ability* for the low/high ability comparison group. In connection with the above-mentioned conditioning, it is also recommended that a study be conducted using Logistic Regression or Mantel-Haenszel Statistic procedure, incorporating more than two or multiple ability estimates in DIF analysis. That is, matching should be conditioned simultaneously on total score, a categorical variable, and additional educational background variables like age, verbal ability, mathematical ability, social class, educational attainment, type of community, and the like.

Future studies should focus on other issues. These include matters related to comparative study of *item response theory* (IRT) and *contingency table* (CT) methods on any relevant psychometric issue, such as of test equating and item banking.

Test experts and developers should pay increasing attention to equity of test scores for various subpopulations of examinees, be it regular or students with learning disabilities. Test equity can be achieved by ensuring that a test measures only construct-relevant differences between subpopulations of examinees. To achieve test equity, bias testing must be conducted especially for very important tests like entrance examinations and professional licensure examinations.

One of the objectives of this paper is to detect DIF and potentially biased items. However, it is also recommended that further studies be conducted to go beyond detecting DIF/biased items and obtain additional information about DIF/biased items. Some items may show larger magnitude of DIF/bias, while some others show relatively small magnitude of DIF/bias. In such a situation, it is of interest to investigate sources of such variation.

## References

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4), Sage Publications, Inc., California.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*(4), *453-464.*

Fraenkel, J. R., & Wallen, N. E. (1994). *How to design and evaluate research in education* (3rd ed.). NJ: McGraw-Hill Book, Inc.

Gierl, M. J. (1999). Differential Item Functioning on the Alberta Education Social Studies 30 diploma examination. *Canadian Social Studies, 33*(2), 66-84.

Kamata, A., & Vaughn, B. (2004). An introduction to Differential Item Functioning analysis. *Learning Disabilities: A Contemporary Journal, 2*(2), 49–69.

Kanjee, A. (2007). Using logistic regression to detect bias when multiple groups are tested. *South African Journal of Psychology, 37,* 47–61.

Mazor, K. E., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect Differential Item Functioning. *Journal of Educational Measurement, 32, 131 – 144.*

O'Neill, K. A., & McPeek W. (1993). *Item and Test Characteristics that are Associated with Differential Item Functioning*. NJ: Educational Testing Service.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting Differential Item Functioning. *Applied Psychological Measurement, 17, 105 – 116.*

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning using logistic regression procedures. *Journal of Educational Measurement, 27, 361 – 370.*

## About the Author

**Jose Q. Pedrajita** is full time faculty member of the Research and Evaluation Area of the College of Education, University of the Philippines, Diliman, Quezon City. He teaches subjects in Tests and Measurement, Research Methods in Education, Statistical Methods Applied to Education, Special Topics in Research and Evaluation, and Research Seminar in Education. He is also part of the College Academic Personnel Committee, Master Admission Test Examination Committee, Doctoral Admission Test Examination Committee, and College Research and Development Committee.

Educational Measurement and Evaluation Review (2011), Vol. 2, 111-123


http://pemea.club.officelive.com/ EMEReview.aspx

# Attributes of Distracters of Multiple Choice Items with Potentially Biased Options

Janet Lynn S. Montemayor
*Benguet State University*
*University of the Philippines-Diliman*

**Abstract** For a test to generate consistent results across subgroups, its components should be objective. This study examined the attributes of the potentially biased options of the Research Competency Test for Graduate Students in Education. The RCTGSE was administered to graduate students in the Masters level (*N*=275) enrolled in private universities (*N*=122, 44%) and in state universities (*N*=153, 56%) from the highlands (*N*=145, 52.7%) and the lowlands (*N*=51, 18.5%). Of the 53 items subjected to distracter response analysis, five were detected with potentially biased options: one against students in government universities, two against students from the highlands, and two against students from the lowlands. Ambiguity and miskeying, and defective distracter are the prominent attributes of the distracters of items with potentially biased options.

**Keywords:** *distracter response analysis, item bias, distracter analysis, research competency*

For legitimate use of tests, it is crucial that all examinees be assured of equity in test items as far as their achievement or ability may be reliably assessed (Osterlind, 1983). Instruments ought to be constructed in such a way that they generate consistent results regardless of affiliation in a subgroup of the testing population. In other words, tests should be bias-free. *Bias* is defined as a systematic error in the measurement process. It affects all measurements in the same way, changing measurement involves sometimes increasing it and other times decreasing it. Items are judged as being relatively more or less difficult for a particular subgroup by comparison with the performance(s) of another subgroup or groups drawn from the same population (Osterlind, 1983).

Test bias reflects psychometric inequalities among subgroups and can take the form of relationship bias or measurement bias. Relationship bias is concerned with the association between a test score and an external criterion measure, while measurement bias is concerned with the properties of test items (Drasgow, 1984 cited in Sheppard et al., 2006). Stated alternatively, when members of different groups (e. g., men vs. women) having identical scores on a given latent construct (e. g., cooperation) have different likelihoods of endorsing an item measuring that construct, the item is said to be functioning differently for the two groups or to be demonstrating DIF (Mitchelson et al., 2009). It is somehow necessary to separate the true differences from the artificial differences which are due only to the testing process (Green, Crone, & Folk, 1989).

The concern of bias in psychological tests is construct validity for items, that is, the extent to which a test item (or set of items) may be said to measure a single, definable theoretical construct or trait. When items have the same construct validity for all examinees in a population, examinees of comparable ability should have the same chances of getting the item correct. In test theory, the chances of an examinee of correctly responding to an item is termed the *probability of success*. A test item is said to be unbiased when the probability for success on the item is the same for equally able examinees of the same population regardless of their subgroup group membership. Bias is not the mere presence of a score difference between two groups (Osterlind, 1983).

Test items are interrogated by the various item bias detection techniques available in an effort to determine whether or not they conform to a given set of psychometric rules in the same way for all persons in a population; for instance, ethnicity and gender (Osterlind, 1983). A method commonly used to examine if a test disadvantages a given group is differential item functioning (DIF) analyses (Camilli & Shepard, 1994). DIF is a set of statistical techniques for examining group differences in responses to an item after controlling for the level of the trait or characteristic that is assessed by the particular measure (e.g., intelligence). Because these individuals are equally intelligent, there should be no difference in the probability of correctly answering the item. When individuals from different groups with the same level of the trait have different probabilities of correctly responding to an item measuring that trait, the item is advantaging the group with the higher probability and DIF is said to be occurring (Scherbaum & Goldstein, 2008). DIF occurs when an item on a test or questionnaire has different measurement properties for one group of people versus another, irrespective of mean differences on the construct (Woods, 2009).

In the context of two-group DIF testing, X is group membership, Y is an item score, and Z is a summed score, used to match the reference (R) and focal (F) groups on the construct (Woods, 2009). In DIF applications, rejection of $H_0$ suggests that even after members of the reference and focal groups are matched on some measure of ability (the stratification variable), they tend to differ in their item scores (Zwick & Thayer, 1996).

Two types of DIF often discussed include uniform and nonuniform. Uniform DIF is defined as a difference between the groups in the probability of a correct response to an item at all ability levels. Nonuniform DIF, sometimes called crossing DIF (CDIF), refers to the case where an item discriminates across the levels of ability differently for the groups (Finch & French, 2007).

There are many methods available for DIF assessment. Distracter response analysis (DRA) is a DIF procedure that examines the incorrect alternatives to a test item—usually termed distracters but sometimes called foils—for differences in patterns of response among different subgroups of a population. The distracter approach focuses attention only on response alternatives; there is no assessment on the item stem (i.e., wording of the question statement) or directions in answering. Nor does distracter response analysis investigate test-taking strategies or techniques, such as preparedness, using an answer sheet rather than responding directly on the test booklet, extraneous noise, or other disturbing conditions. The function of DRA is to determine the significance of the differences among two or more groups' response frequencies in the discrete categories of question distracters. If a significance test reveals that two or more groups distinguished by some criterion as in fact differentially attracted to a test item's distracters, the null hypothesis (of no difference in the group's relative frequencies for distracters) may be

rejected and bias is inferred to be present (Osterlind, 1983). In other words, if different groups prefer different incorrect responses to an item, then the item probably means something different to the different groups. Items that have different meanings to different groups would seem to be biased in a very fundamental sense (Green, Crone & Folk, 1989).

Different factors may explain the biases detected. Green, Folk and Crone (1989) claim that differences may be due to different educational and life experiences. And because tests assess present performance, the test scores may reflect real differences in knowledge, skills, and developed abilities.

It is therefore essential to evaluate the attributes of the distracters of the items with potentially biased options. When established, attributes provide indications as to the probable sources of the significant differences between comparison groups in terms of performance in a particular item. This is done by assessing the pattern of the respondents of the students who did well in the test—that is, the upper 27%. Classical item analysis enumerates three attributes of distracters. *Ambiguity* is observed when there is about an equal number of students choosing the correct answer and a particular distracter, m*iskeying* is detected when there are more students who selected a distracter over the correct answer. *Guessing* is recognized when the responses are distributed almost equally to all options, including the correct answer.

This study intended to link the attributes of the distracters of items detected to be potentially biased. In this study, the researcher used ethnicity and type of university as grouping variables to investigate the differences of scores between subgroups in the Research Competency Test for Graduate Students in Education (RCTGSE). *Ethnicity* refers to the classification according to the ethnic origin of the graduate school student: *highlander* and *lowlander*. *Highlander* pertains to those who were born of parents from any of the many tribes of the Cordillera Administrative Region, whereas *lowlanders* are those who were born of parents from any of the tribes outside the Cordilleras. On the other hand, *type of university* refers to the classification of higher education institutions where the graduate student is enrolled during the time of study. *Private university* pertains to institutions that are owned privately or by owned whereas *state university* pertains to institutions that are owned and governed by the government.

Meanwhile, the RCTGSE is a competency test that the researcher developed with the intent of gauging the graduate students' knowledge on the rudiments of research. The RCTGSE consists of 100 four-option multiple choice items, which are distributed among the following areas: 1)research process, 2)research problem, 3)ethics and research, 4)variables and hypotheses, 5)reviewing the literature, 6)sampling, 7)instrumentation, 8)research design and methodology, 9)descriptive statistics, and 10)inferential statistics.

The faculty of the graduate school shall be the ones to mainly benefit from this study as it will enlighten them about the current state of graduate school students' research competencies according to ethnicity and type of university. With this knowledge, the faculty will be able to adjust teaching strategies and course requirements accordingly such that the students may improve their research competency. This study will also serve as an eye opener to the administrators and curriculum developers of graduate studies. Appropriate actions may be taken to advance the research competencies of the graduate students based on evidence about their current standing. Moreover, graduate school students shall also benefit from the results of this study. As students, they need to be aware of what is expected of them academically and how they are measuring against it and against the performance of other subgroups of the population. Through the results of this study,

the students shall be able to seek ways to boost their research capabilities and therefore improve their academic performance.

The graduate school students' scores in the RCTGSE would somehow reflect what has been learned about research in the course of taking their respective degrees prior to graduate studies. As such, the results of this study shall likewise benefit the curriculum developers, administrators, faculty, and students in the college level. Research may or may not be a requisite course in the different undergraduate programs. Through this study, curriculum developers shall be able to obtain information as to whether to exclude, include, or enhance the research courses in the various undergraduate programs.

Administrators of HEIs shall gain insight regarding the essence of research to the students in their baccalaureate programs. In effect, they would be able to come up with a line up of activities that would suit the needs of their students as regards their research competencies. The faculty may feel the need to integrate research activities in the various undergraduate courses they teach. In that way, the students shall experience research and the endeavour would not be new when they get to the graduate school level.

Through the results of this study, students in the undergraduate level shall gain insight as to what would be expected of them in the graduate level. In this way, the students would be able to prepare themselves earlier for the challenge of independent research.

Finally, results of this study may be linked to previous results of related researches, which could encourage other researchers to do a further study about the research competencies of graduate school students and different subgroups.

## Research Competencies

For any particular research topic, there are three levels of learning outcomes: being familiar with research, having knowledge of research, and being competent at research. Research competency is a description of this third level and specifies the knowledge and skills required. It does not describe how they may be acquired, although it is useful to provide this information in a competency document. However, it should be borne in mind that there are usually many ways by which a specific competency can be acquired (Royal College of Surgeons, 2007).

Specific competencies refer to clusters of cognitive prerequisites that an individual requires in order to be able to perform adequately in a given substantive area (Weinert, 2001 as cited by Mahmud & Zainol, 2008). Specific competencies usually rely on a system of specialized skills and routines stored in memory. These include: extensive mental networks of specialized knowledge, and automatic action routines that must be controlled at a high awareness level and which are acquired through long term learning, experience and deep understanding of the topic. General competencies on the other hand, include concepts such as intelligence, information processing models, meta-competencies, and key-competencies.

Key competencies allow one to perform tasks in a broad scale of contexts (Allen, et al., 2003, as cited in Mahmud & Zainol, 2008). These include planning for problem solving, competent use of media and computer skills, communicative competencies such as command of foreign languages, rhetoric skills and verbal and written presentation skills, and reasoning competencies such as critical thinking and multidimensional evaluation of one's own actions as well as the actions of others.

Research tasks call for specific research competencies. A number of authors enumerate various skills; research institutions and universities also design research competencies frameworks for their respective constituents.

Research competency is evidenced by the ability to demonstrate knowledge in identifying researchable problems, developing research questions and/or research hypotheses, reviewing relevant literature, matching purpose, design and methods, applying appropriate statistical techniques, interpreting results and finally, effectively communicating the research findings (Mahmud & Zainol, 2008).

## Factors Affecting Research Competencies

A student's performance in the graduate school is a product of many factors, such as: previous research experience, previous research supervisor, previous research adviser, time, funding, nature of job, writing ability, and motivation to do research towards obtaining a higher degree or towards being promoted.

It is important for faculty to stress the value of undergraduate research for graduate school preparation and admission and to provide research opportunities that students can complete before the graduate school application process begins. Research experiences provide important preparation for graduate school (Landrum & Nelson, 2002 as cited in Koch, n.d.).

This is why graduate schools value undergraduate research and use it as a criterion for acceptance into graduate programs (Vittengl et al., 2004 as cited in Baltes, et al., 2009). Some graduate schools assess applicants by way of recommendations coming from previous professors; others do it by way of a test.

Further, students who engage in undergraduate research feel better prepared for graduate school (Huss et al., 2002 as cited in Baltes, 2009). Competency in research for graduate students appears to begin with positive experiences in the early research design courses.

It is then important to ensure that the first core research course experience provides the needed support and mastery experiences to enhance research self-efficacy in graduate students (Baltes, 2009). Research courses that bridge prior learning with new applications for and motivation to conduct research may be the road to building research competency in graduate students.

Involvement in research is important for at least four reasons (Koch, n.d.). First, research can help a student determine his or her area of interest in psychology, thereby allowing for a more focused search of graduate programs. Second, working with a faculty member on research can help yield better letters of recommendation. Third, undergraduate research provides an excellent opportunity to enhance several secondary criteria for graduate school admission. Finally, engaging in research helps develop research-based skills that are important for success in graduate school.

Teachers are a major "make or break" factor for students' success in higher education. Previous professors play a part in the students' performance in research in the graduate school. For this very reason, teacher qualification question remains urgent (Lamanauskas, 2008). The majority of the researches both national and international in one way or another reveal a direct link between students' achievements and teachers' competence.

Meanwhile, the motivation and interest towards research dictate the students' success in graduate school. The usual negative associations with research courses have led to diminished amounts of time spent in and effort spent on research courses and projects (Lei, 2008 & Papanastasiou, 2005 as cited in Baltes et al., 2009).

Selecting an issue that is within their capabilities is essential (Gray, 2004). Skills will hopefully develop during the course of the research process, but choosing a topic that requires statistical skills when a student is comfortable with only basic mathematics may be a recipe for disaster. In research, projects that are congruent with both work area and experience (the safe approach) may be chosen or something beyond both their work and current knowledge set. This poses greater risks, but also enhances opportunities for personal development. Moving the project into unfamiliar work area may also provide opportunities for networking amongst new groups of people which can be advantageous for both the project and the graduate student's own professional future (including their future as a researcher).

Students' interest is somehow also linked with the specialization of the professors in the graduate school. It was observed that graduate school applicants often underestimate the importance of identifying potential thesis/dissertation chairs and members that match their applications. Finding the perfect mentor and future dissertation chair is a key factor to consider ("Highlighting Your Research," n.d.). Graduate students should make sure there are professors studying their specific areas of interest (MacNeill, n.d.).

Basic skills related to research are also determinants of students' research performance in the graduate school. Writing and using the computer are two to name a few. Enrolling in graduate programs entail frequent encounters with writing challenges (Richards & Miller, 2005). Graduate school students meet very rigorous writing demands to complete their programs. In terms of writing the research proposal and the manuscript, students write for audiences who have authority over them (Elbow, 2000 as cited in Richards & Miller, 2005).

Another basic skill is using modern technologies that supposedly aid the undertaking of the research process. Nowadays, data analysis is performed in the computer using various softwares, like the Minitab, Statistical Package for the Social Science (SPSS), and Statistical Analysis Software (SAS). However, the question of using the newest information communication technologies remains problematic (Lamanauskas, 2008).

Theoretical knowledge in the graduate school students' field is also important. Theresearch process requires students to engage at some stage with theoretical perspectives (Gray, 2004). Sometimes, this will occur before undertaking the research (deductive approach) and at other times after it (inductive approach). Graduate school students should make sure that their topics are capable of being linked to the appropriate academic theory (Raimond, 2002).

Another factor that affects the research competency of graduate school students is time. Students in graduate school spend little time on research, especially once they secure a faculty position (National Survey of Student Engagement, 2009, as cited in Baltes, 2009).

However, it is a totally different story when research tasks are embedded in the job of the graduate student. As the real world becomes more competitive, complex and uncertain, many people are recognizing the importance and value of research. Hence, research is no longer just the remit of the professional researcher or the university

academic. It is increasingly becoming an integral part of the job specification for many occupations (Gray, 2004).

## Method

### Participants

Master's level students enrolled in the private and public or state universities in the Cordillera Administrative Region and the Ilocos Region were involved in the study. Purposive sampling technique was employed in selecting the subjects, with type of school and ethnicity of graduate students as criteria. A total of $N$=275 graduate students participated in the study with $N$=122 (44%) enrolled in private universities, $N$=153 (56%) in state universities; $N$=145 (52.7%) from the highlands and $N$=51 (18.5%) from the lowlands.

### Instrument

The Research Competency Test for Graduate Students in Education (RCTGSE) consists of a 100-item cognitive test that purports to gauge the research competency of graduate students. The four-option multiple choice items were constructed based on a table of specifications built on the prescribed contents of the course Methods of Research, specifically: (1) research process, (2) research problem, (3) ethics and research, (4) variables and hypotheses, (5) reviewing the literature, (6) sampling, (7) instrumentation, (8) research design and methodology, (9) descriptive statistics, and (10) inferential statistics.

The RCTGSE was content validated by three research professors in various graduate schools based on the test blueprint. Subsequently, the test was administered to a development sample of graduate students ($N$=20) from private and state universities and from highland and lowland origin for reliability testing and item analysis. Scores were used to determine the reliability of the scores and item analysis. Reliability was established at KR20= 0.84. Reliability is determined by estimating the influence of various sources of error. If there is little error, then the reliability is high or strong. If there is much error, the reliability is low or weak (McMillan, 2005). As the reliability coefficient of the RCTGSE is considered high at KR-20 = .84, it can be deduced that the measurement error incurred in using the test is negligible.

Results obtained from the classical item analysis done in all items were $0.00 < p < 0.95$, and $-0.30 < d < 0.70$. Based on the prescribed scale by Pedrajita (2010) and Nava (2010), items subjected for revision or deletion are those whose indices are $0.20 > p > 0.80$ and $0.30 > d > 0.80$. In effect, a total of 47 items were deleted. Table 1 summarizes the items for inclusion and deletion in the final RCTGSE instrument vis-a-vis the table of specification. The table shows that, despite the deletion of items that are functioning poorly based on their difficulty and discrimination indices, the items for inclusion in the RCTGSE are still distributed across the dimensions as prescribed by the table of specifications.

Table 1
*Summary of Items for Inclusion in and Deletion from the RCTGSE*

| Dimension | Prescribed Number of Items* | Number of Deleted Items | Number of Included Items |
|---|---|---|---|
| The Research Process | *11* | *5* | 6 |
| The Research Problem | *10* | *5* | *5* |
| Ethics and Research | *10* | 3 | 7 |
| Variables | *10* | 6 | 4 |
| Reviewing the Literature | *6* | 2 | 4 |
| Sampling | *10* | 2 | 8 |
| Instrumentation | 8 | *5* | 3 |
| Research Design and Methodology | *15* | 7 | 8 |
| Descriptive Statistics | *10* | 4 | 6 |
| Inferential Statistics | *10* | 8 | 2 |
| Total | 100 | 47 | 53 |

* Based on the table of specifications

## Procedure

Permission from the academic heads of selected universities and colleges was sought. Upon approval, proper coordination was done with the respective graduate school professors as regards the schedule of classes where the RCTGSE was administered.

The RCTGSE were given to the graduate students following the standard procedures of test administration. The answer sheets were then scored and responses to the items were recorded and subjected to appropriate statistical analysis.

Item bias was detected using differential item functioning procedures, particularly distracter response analysis with type of school (private or public/state) and ethnicity (highland or lowland) as grouping variables. Classical item distracter analysis was used to determine the attributes of the distracters of the items with potentially biased options.

## Results

### Distracter Response Analyses

The 53 items in the final version of the RCTGSE were subjected to differential item functioning (DIF) procedure, particularly distracter response analysis, to test the items for biases for or against type of university (private/state) and ethnicity (highland/lowland). Total scores obtained in the RCTGSE (k = 100 items) was treated as matching variable. Table 2 is a summary of the DRA for type of university and ethnicity.

**Type of University.** With matched scores, the score of $N$=240 respondents were included in the analysis ($N_1$=112, 46.67% private; $N_2$=128, 53.33% state). DRA was done on all the three distracters of each item. Obtained values in the distracters of 52 of the 53 items included in the analysis are $p > 0.0125$ (1/b $\alpha$ = ¼ *0.05), considering three incorrect

response alternatives and a fourth for double marks or omissions. Thus, for the 52 items, there is no significant difference between private and public school students' relative frequencies for distracters in the RCTGSE.

However, DRA for option B of item number 26, an item under the subtest *Research and Ethics*, resulted to $\chi^2 = 8.702$ ($p < 0.0125$). Therefore, there is a significant difference between private and public school students' relative frequencies for distracter B of item 26 of the RCTGSE. As there are more respondents from the state universities who picked distracter B ($f_1=14 < f_2=34$), it is inferred that the distracter has a potential bias against students from state universities.

**Ethnicity.** Matching the total scores obtained by the respondents in the RCTGSE, a total of $N=90$ students were included in the distracter response analysis ($N_1=45$ highlands; $N_2=45$ from the lowlands).

The distracters of 49 of the 53 items subjected to distracter response analysis resulted to $p > 0.0125$. Thus, for the 49 items, there is no significant difference between highland and lowland students' relative frequencies for distracters in the RCTGSE. The result implies that the distracters of the items in the RCTGSE are functioning well regardless of the ethnicity of examinees.

Analysis bared $p < 0.0125$ for option D of item 16 ($p = 0.004$), option D of item 17 ($p = 0.005$), option C of item39 ($p = 0.011$), and option D of item 94 ($p = 0.010$). The figures indicate that the number of respondents choosing the particular option is significantly higher on one ethnic group compared to the other. Thus, the options are potentially biased. Option D of both items 16 and 17, which are under the subtest *Research Problem*, are potentially biased against highlanders; whereas option C of item 39 under the subtest *Variables* and option D of item 94 under the subtest *Inferential Statistics* are potentially biased against lowlanders.

Table 2
*Summary of Distracter Response Analysis for Type of University and Ethnicity*

| Item Number | Subtest | Distracter | Response | | $\chi^2_0$ | Sig | Disadvantaged Group |
|---|---|---|---|---|---|---|---|
| | | | Group 1 | Group 2 | | | |
| 26 | Ethics & Research | B | 14 | 34 | 8.702 | 0.003 | Public |
| 16 | Research Problem | D | 12 | 4 | 8.517 | 0.004 | Highland |
| 17 | Research Problem | D | 20 | 10 | 8.041 | 0.005 | Highland |
| 39 | Variables | C | 6 | 15 | 6.470 | 0.011 | Lowland |
| 94 | Inferential Statistics | D | 2 | 11 | 6.658 | 0.010 | Lowland |

## Attributes of Distracters of Items where Potential Bias is Detected

Responses of the disadvantaged groups in items where there are detected potentially biased options were recorded (Table 3). Respondents were categorized according to the scores they obtained from the RCTGSE (k=100), where the high group are the top 27% and the low group are the lower 27%.

With f = 0 in the high group responding to options D and B for items 26 and 94,respectively, it can be deduced that the said options are defective as they do not attract respondents to pick them as answers. Distribution of responses in items 16, 17, and 94 are about the same as the number of students who picked the correct answer. Thus, it is deduced that the distracters are ambiguous. In item 39, there are more respondents who picked the distracter than those who chose the correct answer. Thus, miskeying is inferred.

Table 3
*Attributes of the distracters of items with potentially biased options*

| Item | Disadvantaged Group | Options | Responses in the Disadvantaged Subgroup | | Attribute |
|---|---|---|---|---|---|
| | | | High Group | Low Group | |
| 26 | State University | A | 1 | 4 | Defective |
| | | ^B | 5 | 12 | distracter (D) |
| | | *C | 25 | 15 | |
| | | D | 0 | 0 | |
| 16 | Highland | *A | 15 | 6 | Ambiguity |
| | | B | 10 | 9 | |
| | | C | 7 | 7 | |
| | | ^D | 10 | 10 | |
| 17 | Highland | *A | 13 | 12 | Ambiguity |
| | | B | 2 | 5 | |
| | | C | 5 | 1 | |
| | | ^D | 11 | 13 | |
| 39 | Lowland | A | 4 | 3 | Miskeying |
| | | *B | 3 | 3 | |
| | | ^C | 5 | 5 | |
| | | D | 2 | 2 | |
| 94 | Lowland | *A | 5 | 3 | Ambiguity/ |
| | | B | 0 | 0 | Defective |
| | | C | 5 | 5 | distracter (B) |
| | | ^D | 3 | 5 | |

* - correct answer; ^ - potentially biased distracter

## Discussion

Summarizing the distracter response analyses done on RCTGSE (k=53) resulted to one item with an option that is potentially biased against students in state universities, two items with an option that is potentially biased against students from the highlands, and two items with an option that is potentially biased against students from the lowlands.

Differences may be due to different educational and life experiences (Green, Folk, & Crone, 1989). Private and state universities may be guided by certain standards in research instruction, but the way courses are taught is not controlled. Teaching strategies, learning experiences provided, and evaluation procedures may greatly vary, which are manifested in differing performances in cognitive tests like the RCTGSE. The same may be attributed to the results of DRA when ethnicity is used as grouping variable. Students from the highlands and the lowlands differ in terms of experiences and exposure, which might significantly affect the way they respond to items.

Results show that an item in *Ethics in Research* area has put students in state universities to a disadvantage. Two items on *Research Problems* were found to be biased against students from the highlands while one item on *Variables* and *Inferential Statistics* were detected to be biased against students of lowland origin. Judging from the results of the analysis on the attributes of the distracters of the items with potentially biased options, it can be deduced that the differences are due to the subgroups' differences in background and educational exposure.

Ambiguity results when the number of students who choose the correct answer and those who select a particular distracter are relatively the same. This distracter characteristic is common to three of the five items with potentially biased options. For a reason, the disadvantaged groups are more easily attracted to the potentially biased option. This finding implies that students in the disadvantaged groups can hardly distinguish the correct answer from the distracter they have selected.

Meanwhile, one item with a potentially biased option is found to be miskeyed. The concept of miskeying may be interpreted in two ways: 1) that the item is miskeyed, meaning the wrong option was keyed as the correct answer, or 2) that the students in a particular subgroup interpreted the item stem differently, thus selecting a distracter instead of the correct answer.

## Recommendations

Based on the conclusions drawn from the results of the study, the following recommendations are put forward: 1) items with potentially biased options should be subjected for content analysis to identify the sources of bias; 2) items of the RCTGSE should be subjected for bias analysis using other differential item functioning procedures; 3) potential bias of the items of the RCTGSE should be tested with other ethnic groups in the Philippines; and 4) the reliability of RCTGSE should be tested in the doctorate level.

## References


Almack, J.C. (1930). *Research and thesis writing*. USA: The Riverside.

Azuma, R.T. (2003). *So long, and thanks for the PhD!*. Retrieved on September 5, 2009 from http://www.cs.unc.edu/~azuma/hitch4.html

Baltes, B., Hoffman-Kipp P., Lynn, L. and Weltzer-Ward, L. (2009). *Students' research self-efficacy during online doctoral research courses*. Ninth Annual IBER and TLC Conference Proceedings.

Barnes, R. (1995). *Successful study for degrees*. New York, USA: Routledge.

Bieschke, K.J. et al. (1993). *A factor analysis of the research self-efficacy scale*. Retrieved on September 23, 2009 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/90/a8.pdf

Calmorin, L.P. and Calmorin, M.A. (2007). *Research methods and thesis writing* (2nd ed). Quezon City, Philippines: Rex.

Camilli, G, and Shapard, L.A. (1994). *Methods for identifying biased test items.* USA: Sage.

Chon, K.S. (n.d.). *The practice of graduate research in hospitality and tourism*. Retrieved on December 16, 2009 from http://books.google.com.ph/books.

Denscombe, M. (2002). *Ground rules for good research: A 10-point guide for social researchers*. Philadelphia, USA: Open University.

Eaton, K. (2007). *Defining competencies in primary dental care research.* [on-line available] http://www.ingentaconnect.com/content/rcse/brcs/2007.

Edward, J. (2008). *Choosing the dissertation topic for PhD dissertation.* [on-line available] http://www.articlesbase.com/print.

Faculty of General Dental Practice (UK)-The Royal College of Surgeons of England. *Research Competencies Framework.* [on-line available] http://www.fgdp.org.uk/pdf/competencies.pdf.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning. *Educational and Psychological Measurement, 67*(4), 565-582.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education (6th ed).* USA: McGraw Hill.

Garcia, A. M., Nuevo, J. M., & Sapa, E. N. (2007). *Research for all disciplines*. Valenzuela City, Philippines: Mutya

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distracter functioning. *Journal of Educational Measurement, 26*(2), 147-160.

Koch, C. (n.d.). *The value from the graduate school perspective.* [on-line available] *teachpsych.org/resources/e-books/ur2008/7-7%20Koch.pdf.*

Kotrlik, J. W., Bartlett, J. E. II, Higgings, C. C., & Wiliams, H. A. (2002). Factors associated with research productivity of agricultural education faculty. *Journal of Agricultural Education, 43*(3), 23-45.

Lamanauskas, V. (2008). *Some ideas about science and technological education actualities and perspectives.* Problems of Education in the 21st Century (Vol 9).

Lei, S.A. (2008). *Factors Changing Attitudes of Graduate School Students toward an Introductory Research Methodology Course.* [on-line available] http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/ detailmini.

Mahmud, Z., & Zainol, M. S. (2008). Examining postgraduate students' perceived competency in statistical data analysis and their attitudes towards statistics. *International Journal of Education and Information Technologies, 1*(2), 12-20.

Mann, P. S. (2004). *Introductory statistics* (5th ed). Singapore: John Wiley & Sons.

Mithelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2008). Gender and ethnicity differences on the Abridged Big Five Complex (AB5C) of personality traits: A differentiall item functioning analysis. *Educational and Psychological Measurement, 69*(4), 613-635. doi: 10.1177/0013164408323235

Peterson, K. (2004). *Research "strings": their nature, scope and impact* [online site]. Retrieved from http://www.universityofcalifornia.edu/senate/news/source/source2_5a.pdf.

Phillips, E. M., & Pugh, D. S. (2000). *How to get a PhD*. Buckingham, England: Open University.

Richards, J. C., & Miller, S. K. (2005). *Doing academic writing in education.* New Jersey, USA: Laurence Erlbaum.

Rugg, G., & Petre, M. (2004). *The unwritten rules of PhD research.* England: Open University.

Sharp, J. A., Peters, J., & Howard, K. (2002). *The management of a student research project* (3rd ed). England: Gower

Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*(4), 537-553. doi: 10.1177/0013164407310129

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment, 13*(442), 442-453. doi: 10.1177/1073191106289031

Walliman, N. (2005). *Your research project* (2nd ed). London: Sage.

Woods, C. M. (2009). Testing for differential item functioning with measures of partial association. *Applied Psychological Measurement, 88*(7), 538-554. doi: 10.1177/0146621608329506

## About the Author

**Janet Lynn S. Montemayor** is a faculty of the College of Teacher Education of Benguet State University where she also obtained her Bachelor of Secondary Education (Mathematics) and Master of Arts in Education (Educational Administration and Supervision) degrees. After eight years of teaching and administrative service, she was sent by the university to take PhD Educ (Research and Evaluation) on scholarship from the University of the Philippines-Diliman. She is currently on her junior year. Her interests include assessment and item response theory. Further correspondence can be sent to janetlynn_80@yahoo.com.

Educational Measurement and Evaluation Review (2011), Vol. 2, 124-136


http://pemea.club.officelive.com/ EMEReview.aspx

# "We're classmates, can we be friends?": Translation and Validation of the Filipino Version of Classmates' Friendship Questionnaire (CFQ) in the Philippines

**Fraide A. Ganotice, Jr.**
**Jonalyn B. Villarosa**
*Palawan State University -Philippines*

**Abstract** It is assumed that friendly students are those who have a deep concern for and understanding of others. Because of this, a number of teaching strategies are used to promote the development of friendliness among learners. Due to the inherent need to develop this trait among learners, a friendliness scale becomes indispensable. The primary objective of this investigation is to describe the translation process done on the original English version of the Classroom Friendship Questionnaire (Miscenko & Rascevska, 2008) and determine the psychometric acceptability of the newly developed Filipino version (CFQ-F) of the scale. The CFQ-F with four dimensions: social contacts, trust, support and cooperation, and lack of hostility was completed by a sample of 466 high school students. Both within-network and between-network approaches to construct validation were adopted in the study. Results of statistical analyses suggest that the instrument has a good internal consistency and support is provided for its construct validity in terms of its factorial structure and correlations with other external variables. Using confirmatory factor analysis, the authors found strong evidence for the four-factor structure of CFQ-F. The psychometric characteristics of this scale justify its usefulness in future research involving Filipino participants. Implications for cross-cultural research are discussed.

**Keywords:** *classroom friendship questionnaire, CFA, Filipino students, translation, validation*

## Introduction

As teachers, we want our students to be cooperative and friendly. This is built on the assumption that once these traits are internalized by the students, their ability to interact and get along well with others improves and they will be more concerned with one another inside and outside the classrooms. In fact, Rubin, Bukowski, and Parker (2006), suggested that when students can make friends and get along with others, they feel better about themselves, feel more connected to school and their classmates, and have the opportunities to practice and develop social skills. Within the confines of the classroom, these valued character traits of the students are assumed to be developed and/or nurtured by the explicit

use of teaching strategies (e.g., group dynamics) where they are given the chance to interact and/or collaborate with one another and accountability on how other students learn in the class is emphasized (see Eggen & Kauchak, 2010; Woolfolk, 2007). In the past, a number of studies examined the effects of cooperative learning (assumed to be easily facilitated when the students are friendly) on student learning (Johnson & Johnson, 1990, 1991; Johnson, Johnson, & Holubec, 1986; Slavin, 1990) and have become supportive of the arguments of developmental theories which assume that interaction among students around appropriate tasks increases their mastery of concepts (Damon, 1984).

It can not be denied that the classroom is a good avenue to develop friendship among students. This assumption is advanced because a greater amount of time is spent by students in class than at home; hence, fostering friendly relationships can be nurtured (Miscenko & Rascevska, 2008). The challenge then hinges on the part of the teacher on how he/she will design teaching-learning activities that are supportive of and/or in harmony with the development of friendliness among students.

The importance of friendliness as an educational and psychological variable necessitates the development of a measure that will capture the very essence of this construct. The development of such measure signals the richness of theorizing studies which can be conceptualized involving a friendship questionnaire. However, a review of research on friendship did not reveal the existence of a questionnaire suitable for use in research on adolescents in the classroom setting (Mendelson & Aboud, 1999; Aldridge, Fraser, & Huang, 1999; Hunter, Boyle, & Warden, 2006). An exemption to this was the recent scale developed by Miscenko & Rascevska (2008) which is the Classmates' Friendship Questionnaire (CFQ). This scale was designed for Latvian population using Latvian language and was eventually translated into English for wider use.

Adaption of psychological instruments developed within definitive cultural backgrounds is a healthy practice among researchers. However, validation of instrument is indispensable if cross-cultural comparison is to be made before meaningful cross-cultural comparisons can be done (Van de Vijver & Hambleton, 1996; Van de Vijver & Leung, 1997; Fischer, 2004) and it is undeniably an essential phase designed to ascertain the psychometric properties of a foreign-made measure when applied to local setting. While it is an acceptable practice to adopt and/or adapt a foreign-made instrument, caution is necessary on its application if it is to be used with groups other than its intended population (Hambleton, 2001). In the Philippine setting, students' cultural backgrounds vary and these have to be considered when a researcher does cross-cultural investigation. It is therefore necessary for him/her to grapple with the idea associated with the importance of validation of the instruments when it is used within the Philippine classrooms.

CFQ was conceptualized primarily for students in a different culture, hence the need to look into the cultural dimension of the scale. As emphasized by Maneesriwongul and Dixon (2004): "The values reflected by an instrument and the meanings of its component constructs may vary from one culture to another. Research instruments must be reliable and valid in each culture studied" (p. 175). In keeping with this, although it is an acceptable practice to adapt foreign-developed measures for local use, it is important that the issue of validity be addressed before the results of such psychological tests with different cultures are interpreted (Hambleton, 2001; van deVijver & Hambleton, 1996; van de Vijver & Tanzer, 2004). It can be recalled that in some cases, psychological test or scales in English administered to bilingual Filipino participants seem to have the same psychometric properties as Filipino translations of the scales (e.g., Bernardo, 2008a;

Bernardo, Posecion, Reganit, & Rodriguez-Rivera, 2004). However, in other cases, the tests seem to yield very different results when they are presented to Filipino bilinguals (e.g., Watkins & Gerong, 1999). Zhang & Bernardo (2000) even suggested that language may be one of the reasons why learning-related psychological scales are not valid with low-achieving Filipino students.

This two-phase validation study therefore describes the translation process done on the English version of the CFQ (the source language) to Filipino language (the target language). The current study subscribes to a construct validation approach (Marsh, 1997) to the empirical assessment of the structure of the Filipino version of the CFQ. It can be noted that studies that adopt this approach can be classified as either within-network or between-network studies. Within-network construct validation, also called internal construct validation pertains to the analysis of factor structure and factor correlation matrix. On the other hand, between-network construct validation or external construct validation approach entails examining patters of relationships between the scales and other theoretically-related constructs (see Marsh, 1997). Specific to the present study, we used both approaches. First, we conducted a within-network approach using confirmatory factor analysis to test the four-factor structure of the CFQ-F. Consistent with the construct validation approach, it is not only important to address validity within an instrument but it is also imperative to explore the possible differential relationships of the construct being studied to other theoretically-relevant measures. Our purpose for undertaking this study is to develop a Filipino version of the Classroom Friendship Questionnaire that could be used for different types of Filipino students, not only for assessment purposes but for research purposes as well.

## The Original Version of the Classroom Friendship Questionnaire

The authors of the 26-item CFQ English version developed and validated the said instrument involving 264 participants from schools in Latvia. From its conceptualization, the measure was originally comprised of 48 items but after employing psychometric procedures, the items were trimmed down to 26. The authors - Miscenko and Rascevska (2008), performed principal components analysis using varimax rotation and yielded four factors: trust, support and cooperation, social contacts out of school, and lack of hostility. In their study, it was found out that all CFQ scales had high internal consistency and test-retest reliability, and correlated significantly with Peer-Relations subscale. Towards the end of their report, Miscenko and Rascevska (2008) suggested that "it will be helpful to carry out confirmatory factor analysis based on data from a larger and more representative sample." Somehow, the present study is a response to their invitation.

## Cross-cultural Adaptation of Instruments

It is observed that a number of theories and psychological scales or measures that have long dominated the psychological literature are based on Western values and research. With this observation, it may be safe to argue that these measures may not be relevant to non-Western contexts (Bond, 1996; Enriquez, 1993; Markus & Kitaynama, 1991). More specifically, research in educational psychology has shown that Western theorizing on psycho-educational constructs may not be appropriate in non-Western settings including the Chinese (Salili, 1995; Watkins & Biggs, 1996; Yang & Yu, 2007; Yu

& Yang, 2003). The Classroom Friendship Questionnaire is not an exception. The assumption that psychological constructs developed and standardized in one culture are widely universal is not warranted. While it is an acceptable practice to adapt foreign-developed measures, cross-cultural researchers conducting studies with individuals from different cultural groups need to consider whether the scores obtained are comparable. Equivalence and bias are important issues that need to be addressed before meaningful cross-cultural comparisons can be made (Van deVijver & Leung, 1997). As Maneesriwongul and Dixon (2004, p.117) noted:

> "The values reflected by an instrument and the meanings of its component constructs may vary from one culture to another. Research instruments must be reliable and valid in each culture studied. Thus, quality of translation and validation of the translated instrument plays a significant role in ensuring that the results obtained in cross-cultural research are not due to errors in translation, but rather are due to real differences or similarities."

While adaptation of educational and psychological tests is a practice common to researchers, caution should be taken. Researchers should take note of reporting evidence of the reliability and validity of instrument if it is to be used in other cultural settings other than where it was initially developed.

## Overview of the Present Study

The aim of the present study was to evaluate the applicability of the Filipino version of the Classroom Friendship Questionnaire among Filipino high school students. First, we described the translation procedure we adopted in this study. Second, we discussed both within-network and between-network approaches to construct validation adopted in this investigation. Specifically, for the within network test, we conducted a CFA to determine the fit of the data. Further, we also employed a between-network test where we assessed the correlation of the four subscales of CFQ (social contacts, trust, support and cooperation, and lack of hostility) with the subscales of Sense of Self Scale (positive self-concept and negative self-concept).

## Method

### Participants

The participants consisted of 466 Filipino adolescent students ($M$ = 186, $F$ = 280) from first year to fourth year high school in two government-owned institutions. A total of 167(35.83%) participants from first year, 131 (28.11%) participants from second year, 105 (22.53%) participants from third year, and 63 (13.52%) participants from fourth year were recruited for the study. Their ages ranged from 12 to 15 years old ($M$=12.95, $SD$ = 1.43).

## Measures

**Classroom Friendship Questionnaire (CFQ).** The original English version of this questionnaire was developed by Miscenko and Rascevska, (2008). The CFQ is a 26-item measure in which each item denotes one of the three classroom friendship domains. These domains include: social contacts ("*Dinadalaw ako ng mga kaklase ko",* $\alpha$ = .76); support and cooperation ("*Humihingi ako ng tulong mula sa mga kaklase ko*", $\alpha$ = .66), (b) trust ("*Ibinabahagi ko ang aking mga sekreto sa aking mga kaklase*", $\alpha$ = .79), and hostility ("*Inaasar ko ang aking mga kaklase sa pagtawag ng di magagandang bansag o pangalan sa kanila*", $\alpha$ = .61). Items were designed to be rated on a 5-point scale with responses ranging from 1 (never/*hinding-hindi*) to 5 (always/*madalas*).

> The Filipino version of the CFQ was given with the following directions:
> "Basahin at unawaing mabuti ang bawat aytem na may kinalaman sa inyong relasyon sa bawat isa sa classroom. Bilugan ang bilang na nagpapakita ng antas ng inyong pinakatapat na sagot."

**Sense of Self Scale (SoS).** Two subscales of SoS namely, positive and negative self-concept drawn from the Inventory of School Motivation (Maehr & Braskamp, 1986; McInerney, Roche, McInerney, & Marsh, 1997), were included specifically to establish the between-network construct validity. More specifically, we would like to note that the Filipino version of the SoS (see Ganotice, 2010; Ganotice, Bernardo, & King, in press) was used in this investigation. The said Filipino version has been validated through confirmatory factor analysis involving Filipino adolescent students.

## Translation Procedure

The Filipino version of the CFQ was developed by using the forward and back-translation methods (Fischer, 2004). It can be noted that back translation is commonly used and regarded as a standard method for translating item wordings of a scale from one language to another, and this method has been recommended by a number of scholars (see Behling & Law, 2000; Chang, Chau, & Holroyd, 2003; Hyrkas, Appelquist-Schmidlechner, & Paunonen-Ilmonen, 2003 for reviews) as it gives an investigator control over the original instrument and its translation. Back translation was maintained through the procedure described by Brislin's classic back-translation model (1980).

The original English version of the CFQ was first translated into conversational Filipino by the second author of the current study. We would like to note that the translation we adapted in this study was in accordance with Brislin's (1980) framework for pragmatic translation of psychometric instruments. Specifically, all the items in the questionnaire were translated from English (source language) into Filipino (target language) and then transcribed back to English, the source language being the check for consistency. Next, the two versions (original and back translation) were analyzed and compared to ensure the linguistic and conceptual accuracy of contents. Finally, we sought the experts' opinion by presenting to them the original and derived versions for their comments and suggestions. These experts have wide exposure to instruments translations.

Because the CFQ-F was intended for use by Filipino high school students, the Filipino translation involved conversational Filipino, which typically involves code-mixing and borrowings from English. Thus, some items in the Filipino version actually include words in English that are more commonly used by Filipino high school students compared to the Filipino equivalent. In such cases, the borrowed English terms are marked by quotation marks, as shown in the following examples:

(a) Nag-oorganize ako ng "meetings" kasama ang mga kaibigan ko tuwing bakasyon (I have organized meetings with my friends during summer vacation)
(b) Sumasali ako sa "party" ng klase. (I have taken part in class parties).

## Administration

The CFQ-F was administered to participants in class groups by the first author, with the assistance of student-teachers assigned in each classroom. To standardize the administration of CFQ-F across class groups, student-teachers who assisted in the administration of the said measure received a copy of the instrument, along with written instructions. Further, the researchers verbally briefed the participating student-teachers about the structure, purpose, and administration of the CFQ-F, prior to its administration with participants. In particular, student-teachers were instructed not to interpret any of the CFQ-F items for the students, but to instruct them to leave an item out or use their best judgment if they did not understand it.

## Statistical Analyses

Consistent with the objectives set for this investigation, we first looked into the descriptive statistics of the CFQ-F and its Cronbach's alpha as a measure of internal consistency reliability. As a test of between-network construct validity, we looked into the correlations of the CFQ-F dimensions with the dimensions of SOS. Finally, we conducted confirmatory factor analysis (CFA) to determine whether the hypothesized structure of the CFQ which Miscenko and Rascevska (2008) proposed was applicable to the Filipino cultural landscape. Confirmatory factor analysis assesses the extent to which items reflect the underlying constructs. Model fit is assessed by a combination of model fit indices. In this study, chi-square statistic and other goodness-of-fit indices such as the Goodness-of-Fit Index (GFI), Adjusted Goodness-of-Fit Index (AGFI), incremental fit index (IFI), Tucker-Lewis Index (TLI), comparative fit index (CFI), root mean square error of approximation (RMSEA), and chi-square/degrees of freedom ratio were used. It is generally accepted that in good measurement models, the GFI, AGFI, IFI, TLI, and CFI will be above .90 while the RMSEA will be below .08 (Byrne, 2001). The chi-square/degrees of freedom ratio should also be non-significant. However, researchers have found that this is usually overly sensitive to sample size differences (Anderson & Gerbing, 1988; Huang & Michael, 2000). Thus we decided to focus on the other fit indices to assess the fit of the hypothesized model.

## Results

### Preliminary Analysis

Preliminary analysis was performed to check the properties of the data. No outliers were found; mean value replacement was used for the few missing data points. The evaluation of the assumptions about multivariate normality and linearity were all satisfactory.

In Table 1, we reflect the means, standard deviations, and internal consistency coefficients for each of the analyzed variables. Acceptable thresholds for internal consistency reliability (alpha) are typically set at .70 (Nunnally, 1978) or .80 (Henson, 2001). Specific to the present investigation, the Cronbach's alphas obtained ranged from .61 - .79. Results indicated that the CFQ-F had an acceptable reliability involving high school Filipino students in the Philippines.

Table 1
*Descriptive statistics, reliability indices, and intercorrelation among variables of the Classroom Friendship Questionnaire (CFQ-F) and Sense of Self Scale (SoS)*

| *Classroom Friendship Questionnaire* | Range | Alpha (α) | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Social contacts of out of school | 1-5 | .76 | 3.27 | 0.74 | - | | | | | |
| 2. Trust | 1-5 | .66 | 3.20 | 0.83 | **0.61*** | - | | | | |
| 3. Support and cooperation | 1-5 | .79 | 3.71 | 0.53 | **0.56*** | **0.52*** | - | | | |
| 4. Lack of hostility | 1-5 | .61 | 2.72 | 0.61 | 0.08 | **0.15*** | 0.02 | - | | |
| 5. Negative self-concept | 1-5 | .78 | 1.86 | 0.64 | -0.03 | **-0.16*** | -0.02 | 0.07 | - | |
| 6. Positive self-concept | 1-5 | .79 | 4.56 | 0.42 | 0.01 | **0.18*** | -0.00 | -0.06 | **-0.52*** | - |

*$p$ <.05;
*Note*. higher scores indicated the greater endorsement of the item

It can be noted that a number of subscales of CFQ are correlated to one another. In the same way, two of the subscales of the Sense of Self Scale (SOS) were significantly correlated with subscales of the CFQ (i.e., negative self-concept and trust; positive self-concept and trust). This result is within expected direction where it can be assumed that individuals who are high in the positive self-concept construct trust others. Support and cooperation and social contacts out of school did not significantly relate with SOS constructs (negative and positive self-concept) perhaps because having high and positive self-concept does not provide a strong theoretical assumption of the possibility that students will create out of school social contact.

## Confirmatory Factor Analysis

In order to realize the objectives set for this study, we conducted CFA to see whether the postulated four-factor structure is tenable involving Filipino participants. The CFA results from the Filipino sample supported the hypothesized four-factor model with a good fit to the data (refer to Table 2). The CFA yielded an excellent fit to the data. Further, the standardized factor loadings were all significant.

Table 2
*Summary of the Goodness-of-Fit Statistics*

| Measure | $\chi^2$ | *df* | $\chi^2/df$ | RMSEA | NFI | NNFI | CFI | GFI | AGFI |
|---|---|---|---|---|---|---|---|---|---|
| CFQ-F | 27.78 | 14 | 1.26 | 0.01 | 0.99 | 0.99 | 0.99 | 0.97 | 0.96 |

*Note. df* = degrees of freedom; RMSEA = root mean square error approximation; NFI = normed fit index, NNFI = non-normed fit index; CFI = comparative fit index; GFI = goodness-of-fit index; AGFI = adjusted goodness-of-fit index.

The results showed good fit for between the data from the Filipino version and the hypothesized four-factor structure of the CFQ. The results of the CFA show very strong support for the four-factor structure of the CFQ-F.

## Between Construct Network Validity

To establish the between-network construct validity, we examined the relationship of the CFQ-F with specific dimensions of the Sense of Self Scale (positive sense-concept and negative self-concept). In the earlier study of Miscenko and Rascevska, (2008) they performed between-network construct validity by correlating specific dimensions of the CFQ with the Peer-Relations Sub-scale of the Self-esteem questionnaire (Hunter, Boyle, & Warden, 2006) specifically for convergent and divergent validity. Specific to the present study, we speculated that *trust* - one of the specific dimensions of the CFQ would be positively correlated with positive self-concept, but negatively associated with negative self-concept. The basic assumption here lies on the nature of the constructs. That is, on one hand, students who trust and serve as channels of support and cooperation with others will be most likely have positive self-concept. On the other hand, those who do not trust others and do not show support and cooperation with others might be manifesting negative self-concept.

Results indicate that scores on the CFQ-F are significantly correlated with positive self-concept and negatively correlated with negative self-concept. These results are within expected direction. The correlations of the specific subscales of the CFQ-F with the SOS subscales provide evidence for the between-network validity of the Filipino translation of the CFQ.

## Discussion

In this research, we set out to examine the factor structures of scores on the 26–item Classmates' Friendship Questionnaire (CFQ) using a large sample of Filipino high school students in the Philippines. We also wished to examine both the within-network and between-network construct validity of the Filipino version of the CFQ.

The results, with the 466 sample Filipino high school students, suggest that confirmatory factor analysis conducted for the ISM scales indicates a good fit between the models and the data using an array of goodness-of-fit indices. Thus, the CFA approach used in this study provided a strong validation, with the Filipino sample, of the Classroom Friendship Questionnaire-Filipino version. Specifically, the data were a good fit for the four-factor model of the CFQ-F.

The internal reliability of the factors which composed the CFQ-F were all adequate reaching acceptable levels, with Cronbach alpha values meeting the criterion. These results provided us with confidence that these instruments may yield valid scores in the Philippine setting even if this was designed for Latvian population. Perhaps this echoes the possibility of friendship being a construct that people from different cultures have in common (etic) which is extensively discussed within the framework of etic-emic model (see Davidson, Jaccard, Triandis, Morales, & Diaz-Guererro, 1976). In other words, while cultures may differ qualitatively in terms of the different meanings assigned to friendship, it seems that this construct has universal and/or normative appeal and was interpreted in a similar fashion by the Latvian and Filipino participants.

The findings of the present investigation hold substantial conceptual and methodological implications for researchers studying issues relevant to friendship in the Philippine classrooms. *First*, on these consistent statistical evidences, validation of the CFQ-F is imperative for educational psychologists, teachers, and school administrators whose aim is to improve classroom learning and to develop genuine concern and interest between and among learners. The 26-item Filipino version of the CFQ may provide a means by which researchers and teachers can attain deep and profound understanding of important dimensions of students' friendship within the confine of the classrooms. In keeping with the finding of Antil, Jenkins, Wayne, and Valdasy (1998, p. 254) where they noted that "nearly 90 percent of elementary teachers used some form of cooperative learning as part of their instruction in schools today," somehow the use of the CFQ-F can be a good starting point worth considering. To do away with the routine, perhaps teachers can consider administering the CFQ-F to have a feel on how close the students are and the results can be initialized as they design group dynamics related to instructional objectives. *Second*, we feel that the CFQ-F deserves to form part of teachers' hosts of assessment tools because its constructs were found applicable to Filipino students. Specific but not limited to guidance counselors and psychologists, perhaps the instrument can be utilized in developing a complete profile of our students which can be an additional basis for understanding their individuality.

Conversely, this research has some important limitations. First, we acknowledge that the sample size though large was not taken from diverse Filipino adolescent populations and thus only represents the government-owned institutions. Perhaps other researchers can consider a wider sample from different high schools including private and sectarian secondary schools. Second, we only tested for the correlation of the CFQ-F with positive and negative self-concepts taken from the SoS scale. Future research could test the relationship of the said scale with a wider range of psycho-educational constructs.

In spite of these limitations, however, we are still positive that the internal virtue of this study remains strong. Overall, the results of this research support the internal consistency reliability and construct validity of the CFQ-F. An advantage of the present study was the use of both within-network and between-network approaches to construct validation. This suggests a stronger case for the validity of the instrument.

In sum, the completion of this study further promotes the importance of the translation/adaptation of existing tests and instruments for local use (Hambleton, Merenda, & Spielberger, 2005) which we started doing in the past (see Ganotice, 2010; Ganotice, Bernardo, & King, in press; Ganotice, 2010a) where we translated foreign instruments and checked their acceptability on the bases of various statistical analyses. It can be recalled that this issue has been extensively addressed in the cross-cultural psychology research community (see Hambleton, 2001; van de Vijver & Hambleton, 1996), but might not be given as much attention in psychology research with Filipino students because the students are assumed to be bilinguals proficient in English, since the medium of instruction in most Philippine schools is English. Perhaps, it is an unwarranted assumption if we think that English psychological scales will be valid for use among Filipino students, even if they are studying in learning institutions where the medium of instruction is English. Thus, we encourage more translations and validations of translations of psychological tests in Filipino and other Philippine languages, and ideally the development of more indigenous psychological tests in the local languages.

As a final note, in recent International Testing Commission guidelines for translation and adapting tests (2010, p. 20), it was recognized that "the growing recognition of multiculturalism has raised awareness of the need to provide for multiple language versions of tests and instruments intended for use within a single national context." We believe that our small effort to translate and/or validate the CFQ-F for use with Filipino-English bilingual students in the Philippines is in harmony with and responsive to this challenge.

## References


Aldridge, J. M., Fraser, B. J., & Huang, T-Ch. I. (1999). Investigating classroom environment in Taiwan and Australia with multiple research methods. *Journal of Educational Research, 93*(1), 48-61.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411-423.

Antil, L., Jenkins, J., Wayne, S., & Valdasy, P. (1998). Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice. *American Educational Research Journal, 35*(3), 419-454.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions.* Thousand Oaks, CA: Sage.

Bernardo, A. B. I. (2008a). Exploring epistemological beliefs of bilingual Filipino preservice teachers in the Filipino and English languages. *The Journal of Psychology, 142*, 193-208.

Bernardo, A. B. I., Posecion, O. T., Reganit, A. R., & Rodriguez-Rivera, E. (2004). Adapting the Social Axioms Survey for Philippine research: Validating Filipino and English versions. *Philippine Journal of Psychology, 38*(2), 77-100.

Byrne, B. M. (2001). *Structural equation modelling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Brislin, R. W. (1980). Translation and content analysis of oral and written materials. In H. Triandis & J. W. Berry (Eds.), Handbook of cross-cultural psychology, Vol. 2 (pp. 389-444). Boston: Allyn & Bacon.

Bond, M. H. (1996). *The handbook of Chinese psychology.* Hong Kong: Oxford University Press.

Chang, A., Chau, J., & Holroyd, E. (2003). Translation of questionnaires and issues of equivalence. *Journal of Advanced Nursing, 29,* 316-322.

Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology, 5,* 331-343.

Davidson, A. R., Jaccard, J. J., Triandis, H. C., Morales, M. L., & Diaz-Guererro, R. (1976). Cross-cultural model testing: Toward a solution of the etic-emic dilemma. *International Journal of Psychology, 11,* 1-3.

Eggen. P., & Kauchak, D. (2010). *Educational Psychology: Windows on Classrooms*. Merrill, Upper Saddle River, New Jersey Columbus, Ohio.

Enriquez, V. G. (1993). Developing a Filipino psychology. In U. Kim & J. W. Berry (Eds.), *Indigenous Psychologies* (pp. 152-169). London: Sage.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of Score Adjustment Procedures and Review of Research in JCCP. *Journal of Cross-Cultural Psychology, 35,* 263-282.

Ganotice, F. A. (2010). Confirmatory factor analyses of scores in Inventory of School Motivation (ISM), Sense of Self Scale, and Facilitating Conditions Questionnaire (FCQ): A study using Philippine sample. *The Educational Measurement and Evaluation Review, 1*, 59-77.

Ganotice, F. A., & Bernardo, A. B. I. (2010a). Validating the factors of the English and Filipino versions of the Sense of Self Scale. *Philippine Journal of Psychology, 43*, 81-99.

Ganotice, F. A., Bernardo, A. B. I., & King, R. B. (2010). Validating the two language versions of the Inventory of School Motivation among Filipino bilingual students. Paper submitted to *Journal of Psychoeducational Assessment.*

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17,* 164-172.

Hambleton, R. K., Merenda, P., & Spielberger, C. (Eds.). (2005). Adapting educational and psychological tests for cross-cultural assessment. Hillsdale, NJ: Lawrence S. Erlbaum Publishers.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology Research. *Counseling Psychologist, 34*, 601-629.

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

Hunter, S., Boyle, J., & Warden, D. (2006). Long-term stability and reliability of scores on the peer-relations subscale of the Self-esteem Questionnaire. *Educational and Psychological Measurement, 66*(2), 331-341.

Hyrkas, K., Appelquist-Schmidlechner, K., & Paunonen-Ilmonen, M. (2003). Translating and validating the Finnish version of the Manchester Clinical Supervision Scale. *Scandinavian Journal of Caring Sciences, 17,* 358-364.

International Test Commission (2010). *International Test Commission Guidelines for Translating and Adapting Tests* [on-line site]. Retrieved from http://www.intestcom.org

Johnson, D. W., & Johnson, R. T. (1990). Social skills for successful group work. *Educational Leadership, 47,* 29-33.

Johnson, D. W., & Johnson, R. T. (1991). *Learning together and alone: Cooperative, competitive, and individualistic.* Third Edition. Englewood Cliffs, NJ: Prentice Hall.

Johnson, D. W., Johnson, R. T., & Holubec, E. J. (1986). *Circles of learning: Cooperation in the classroom.* Edina, MN: Interaction Book Company.

Maehr, M. L., & Braskamp, L. A. (1986). *The motivation factor: A theory of personal investment.* Lexington, MA: Lexington Press.

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: A methods review. *Journal of Advanced Nursing*, *48*, 175-186.

Marsh, H. W. (1997). The measurement of physical self-concept: A construct validation approach. In K. Fox (Ed.), *The physical self-concept: From motivation to well-being* (pp. 27–58). Champaign, IL: Human Kinetics

Markus, H. R., & Kitayama, S. (1991). Culture and self: Implications for cognition, emotion and motivation. *Psychological Review, 98*, 224-253.

Mendelson, M. J., & Aboud, F, E. (1999). Measuring friendship quality in late adolescents and young adults; McGill Friendship Questionnaires. *Canadian Journal of Behavioural Science/Revue canadienne des scienes du comportement, 31,* 130-132.

McInerney, D. M. (2008). Personal investment, culture and learning: Insights into School Achievement across Anglo, Aboriginal, Asian, and Lebanese students in Australia. *International Journal of Psychology, 43,* 870-879.

McInerney, D. M., Roche, L. A., McInerney, V., & Marsch, H. W. (1997). Cultural perspectives on school motivation. *American Educational Research Journal, 34,* 207-236.

Miscenko, T., & Rascevska, M. (2008). Psychometric properties of Classroom Friendship Questionnaire. *Baltic Journal of Psychology, 9,* 129-140.

Nunnally, J. C. (1978). Introduction to psychological measurement. New York: McGraw-Hill.

Rubin, K., Bukowski, W., & Parker, J. (2006). Peer interactions, relationships and groups. In N. Eisenberg (Vol Ed.), *Handbook of child psychology: Vol. 3. Social, emotional and personality development* (pp. 571-645). Hoboken, NJ: John Woley & Sons.

Salili, F. (1995). Explaining Chinese motivation and achievement. In M.L. Maehr & P.R. Pintrich (Eds.), *Advances in motivation and achievement: Culture, motivation, and achievement* (pp. 73-118). Greenwich, CT: JAI.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice.* New Jersey: Prentice Hall.

van de Vijver, F., & Hambleton, R. K. (1996). Translating test: Some practical guidelines. *European Psychologist, 1,* 89-99.

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis of comparative research.* Thousand Oaks, CA: Sage.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54,* 119-135.

Watkins, D. A., & Biggs, J. B. (Eds.) (1996) *The Chinese Learner: cultural, psychological, and contextual influences*, Hong Kong/Melbourne: Comparative Education Research Centre/Australian Council for Educational Research.

Watkins, D., & Gerong, A. (1999). Language of response and the spontaneous self-concept: A test of the cultural accommodation hypothesis. *Journal of Cross-Cultural Psychology, 30*, 115-121.

Woolfolk, A. (2007). *Educational psychology* (10th ed.). NJ: Pearson.

Yang, L. Z., & Wang, J. Y. (2007). A Follow-up Study of Self-imposed Delay of Gratification at Age 4 as a Predictor of Children's School-based Social Competences at Age 9. *Acta Psychologica Sinica, 39*, 668-678.

Yang, L. Z., Xu, L. M., & Wang, J. Y. (2003). A research on 3 to 5 year old children's self-imposed delay of gratification under four attention conditions. *Psychological Development and Education, 4*, 1-6.

Zhang, L. F., & Bernardo, A. B. I. (2000). Validity of the Learning Process Questionnaire with students of lower academic attainment. *Psychological Reports, 87*, 284-290.

## About the Authors

**Fraide A. Ganotice, Jr.** is the chairperson of the Graduate Education Department of the College of Teacher Education, Palawan State University in Puerto Princesa City, Palawan. He obtained his PhD in Educational Psychology (Measurement and Evaluation) from De La Salle University-Manila in June 2010 under the faculty development scholarship of Commission on Higher Education (CHED).

**Jonalyn B. Villarosa** is the chairperson of Secondary Education Department of the College of Teacher Education, Palawan State University in Puerto Princesa City, Palawan. She obtained her MA in Literature from the same institution.

Corresponding Author: Dr Fraide A Ganotice, Jr
Graduate Education Department
College of Teacher Education, Palawan State University
5300 Puerto Princesa City
Palawan, Philippines
email: fraideganotc@yahoo.com
phone: +632-4343732

Educational Measurement and Evaluation Review (2011), Vol. 2, 137-147


http://pemea.club.officelive.com/ EMEReview.aspx

# Exploratory and Confirmatory Factor Analysis of Self-efficacy among Student-Athletes

**Maria Cristina M. Firmante**
*De La Salle University-Manila*

**Abstract** The study developed an instrument that measures four factors of self-efficacy among student-athletes which were based on Bandura's sources of self-efficacy theory. The four factors are: Performance accomplishments, modeling, verbal persuasion, and emotional arousal. The survey questionnaire was validated and distributed to 157 student athletes (*N*= 157). Out of 157 respondents, 87 were males and 70 were females whose ages range from 16 to 21 years old. The respondents were all from one private tertiary institution in Metro Manila. The instrument was tested using Exploratory Factory Analysis (EFA) and the reliability of items, Cronbach's Alpha was also established with an index of .88 that shows good reliability. Out of 66 items, 31 remained significant and stable in the four factors of self-efficacy with a consistency value Cronbach's Alpha of .82. Given the results for the acceptable items in EFA, it was tested using a Confirmatory Factor Analysis (CFA). The goodness of fit based on the RMS standardized residual (RMS=0.070) showed less error having a value closer to .01. The Noncentrality fit indices values shows good fit for self-efficacy with four factor (Steiger-Lind RMSEA = 0.60, Population Gamma Index= 0.910, Adjusted Population Gamma Index= 0.896).

**Keywords:** *Self-efficacy, student-athletes*

## Introduction

Student-athletes in a university face different challenges not experienced by the ordinary college students. Aside from doing their assignments, projects, attending their classes and other extra-curricular activities in and out of school as well as socialization, student-athletes need time to practice well and become more competitive in their sport. These challenges develop efficacious characteristics to help them handle demands in their environment. Self-efficacy is a belief of one's ability and capacity to accomplish or deal with challenges in life. This is the reason why the researcher decided to construct a test in self-efficacy among student-athletes and evaluate the sources of self-efficacy in sports.

Self-efficacy is a very important aspect in dealing with different challenges in life. This would be a good instrument for counselors who handle student-athletes. The results of the test would also help the counselors help the student-athletes succeed in college and in their respective sports. In addition, it can be a basis for developing programs and modules for the student-athletes.

Bandura (1986, 1977) formulated a clear and useful conceptual model of self-efficacy that brought together the concepts of confidence and expectations. The theory of

self-efficacy is the most extensively used theory investigating self-confidence as a sport and motor performance settings (Weinberg & Gould, 1999, 1995). Bandura's (2001, 1997, 1977) self-efficacy theory developed within the social cognitive framework theory where individuals are viewed as proactive agents in the regulation of their cognition, motivation, actions, and emotions rather than as passive reactors to their environment (Feltz, Short, & Sullivan, 2008). In order for self- efficacy to develop, the individual must believe in control and perform intentionally. The power and will to start a course of action is the key feature of personal activity. If a person believes in control and the power to produce specific results, he will be motivated to try to make things happen (Cox, 2002). Bandura (1995) also described self-efficacy as *"the belief in one's capabilities to organize and execute the courses of action required to manage prospective situations."* In other words, self-efficacy is a person's belief in his or her ability to succeed in a particular situation. Bandura (1994) described these beliefs as determinants of how people think, behave, and feel.

In the study of Chu and Tingson (2009) self- efficacy plays an important role in the success and the performance of an athlete. Self-efficacy is commonly viewed as a situation specific variation of self-confidence and has repeatedly been found to be positively related to sporting performance (Moritz, Feltz, Fahrbach, & Mack, 2000). Magno and Lajom (2008) reported that individuals have a sense of self-confidence regarding performance of specific tasks or self-efficacy for learning. Self-efficacy can be influenced by factors such as student abilities, prior experiences and attitude towards learning, as well as by instructional and social factors (Bandura, 1986, 1977; Chu 2011; Cintura, Okol, & Ong, 2001; Jinks & Morgan; 1999; Narciss 2004; Schunk & Cox, 1986).

Self-efficacy influences behaviors including behavioral choice, performance, efforts despite setbacks or recent failures, strategy choice, goal choice, and goal commitment. Research has shown that self-efficacy is a significant predictor of athletic performance; athletes with high level of self-efficacy were found to perform better than those athletes who demonstrated lower levels of self-efficacy prior to competition. These findings were explained by the fact that athletes high in self-efficacy found the competitive situations less threatening and displayed less anxiety than their opponents who were lower in self-efficacy. (Mills, Munroe, & Hall 2000-2001).

According to Vargas-Tonsing, Warners, and Felts (2003) many researches supported Bandura's theory that individual perceptions of self-efficacy can impact a subsequent outcome or performance which is influenced by its four sources: performance accomplishments, vicarious experience (modeling), verbal persuasion and emotional arousal (Bandura, 1977).

Ayiku's (2005) study described student-athletes who have a very different college experience from their non-athlete counterparts (Watt & Moore III, 1993). In addition to attending classes, doing homework, socializing with peers and faculty members, student-athletes must also practice and learn game playbooks while training and performing in their respective athletic endeavors. Athletes may face many challenges to succeed as intercollegiate athletes and as students at institutions of higher learning (Carodine, Almond, & Gratto, 2001; Etzel, Ferrante, & Pinkney, 1996; Ferrante & Etzel, 1991; Howard-Hamilton & Sina, 2001). Athletic or sports self- efficacy refers to the athlete's belief that he or she will be able to proficiently acquire skills of their position(s) necessary to successfully perform at the peak of their athletic performance. It is also concerned with an athlete's belief in his or her ability to achieve personal and team goals which may include everything

from making good snap decisions, to successfully performing learned skills under pressure. (Ayiku, 2005)

The present study further analyzed the factors of self-efficacy among student-athletes using the Exploratory Factor Analysis, after getting the result and most significant and acceptable factors, it was extracted in the Confirmatory Factory Analysis.

## Performance Accomplishments

Performance accomplishments (particularly clear success or failure) provide the most dependable basis for self-efficacy judgments because they are based on one's mastery experiences. If experiences are generally successful, they will raise the level of self-efficacy. However, repeated failures will result in expectations of lower efficacy. For example, if a field goal kicker has kicked the winning field goal in several games as time was running out, he will have a high degree of self-efficacy that he can do it again. Similarly, an athlete rehabilitating from a wrist injury will persist in exercise after seeing steady improvement in her range of motion and wrist strength. Research into diving and gymnastics shows that performance accomplishments increase self-efficacy, which in turn increases subsequent adherence (McAuley, 1985) as well as exercise adherence (McAuley, 1993, 1992;). (Weinberg & Gould 1995, 1999 p. 294).

Bandura (1997, 1986, 1982) said that athletes must experience success in order for self-efficacy to develop. With difficult tasks, this is an unrealistic expectation, so the coach or teacher must ensure success by initially reducing the difficulty of task. The teacher must find a way for beginners to find success, or they will come to believe that they cannot succeed, and quit trying. The difficulty of the task can be increased as the simpler tasks are mastered (Cox, 2002 pp. 19-20).

## Modeling

Physical educators, exercise leaders, athletic trainers, and coaches all often use vicarious experiences also known as demonstration or modeling, to help students learn new skills. This can be a particularly important source of self-efficacy information for performers lacking experience with the task at hand, relying on others to judge their own capabilities. For example, seeing a team member complete a difficult move on the uneven parallel bars can reduce anxiety and help convince other gymnasts that they too can accomplish this move. Bandura (1974, 1965) and McCullagh, Weiss, & Ross (1989), modeling can be best understood through a four-stage process: attention, retention, motor reproduction, and motivation. In order to learn through watching, careful attention must first be given to the model. Our ability to attend depends on respect for the person observed, interest in the capability, and how well can see and hear. The best teachers and coaches do not overload you with information, expect you to focus your attention on all the specific elements of the skill, or show the skill only one quick time. Rather they focus on a few key points, demonstrate several times, and let you know exactly what to look for (Weinberg & Gould, 1999, 1995 pp. 294-295).

Bandura (1997, 1986, 1982) said that in learning new skills the learner needs a template or model to copy. The instructor, a skilled teammate, or a film or video of a skilled performer can provide this. The component of Bandura's theory is the concept of participatory modeling. In participatory modeling, the learner first observes a model

perform a task. Then the model or instructor assists the subject in successfully performing the tasks (Cox, 2002 pp. 19-20).

### Verbal Persuasion

Coaches, teachers, and peers often use persuasive techniques to influence behavior. An example would be a baseball coach telling a player, "I know you're a good hitter, so just hang in there and take your swings. The base hits will eventually come." Similarly, an exercise leader might tell an exercise participants to "hang in there and don't get discouraged, even if you have to miss a couple of days." This type of encouragement is important to participants and can be helpful in improving self-efficacy. When a psychological barrier is present, coaches and instructors sometimes even resort to deception to persuade performers that they can perform certain skills (Weinberg & Gould, 1999, 1995 p. 296).

Bandura (1997, 1986, 1982) found that helpful verbal statements which suggest that the athlete is competent and can succeed are most desirable. Negative comments should always be avoided. Coaching tips can be given in such a way that they do not convey negativism (Cox, 2002 pp. 19-20).

### Emotional Arousal

Although physiological cues are important components of emotions, emotional experiences are not simply the products of physiological arousal. Thus, emotions or moods can be an additional source of information about self-efficacy. For example: an injured athlete who is feeling depressed and anxious about his rehabilitation would probably have lowered feelings of self-efficacy. Conversely, an athlete who feels energized and positive would probably have enhanced feelings of self-efficacy (Weinberg & Gould, 1999, 1995 p. 298).

Bandura (1997, 1986, 1982) argued that proper attention is important in helping the athlete to master a particular skill and develop a feeling of efficacy (Cox, 2002 pp.19-20).

The present study constructed a measure of self-efficacy among student-athletes patterned from Bandura's four sources of self-efficacy; performance accomplishment, modeling, verbal persuasion and emotional arousal (Weinberg & Gould, 1995 pp. 294-298). It was tested using exploratory factor analysis (EFA) to get the most significant and acceptable items and further tested through confirmatory factor analysis (CFA).

## Method

### Item Writing

Items for the Self-Efficacy Inventory (SEI) were constructed based on the sources of Bandura's principal sources of self-efficacy (Weinberg & Gould 1999). The items were classified according to performance accomplishments, modeling, verbal persuasion, and emotional arousal. The scaling technique used was a four-point verbal frequency scale. Each interval in the scale is coded with numerical value where (4= strongly agree; 3= agree; 2= disagree and 1= strongly disagree). There were 100 items judged by experts as to

whether it was accepted, needs revisions, or rejected. Out of 100 items a total of 66 items remained and used in the study. The items were arranged by its category and the respondents answered by encircling the number corresponding to their answer.

## Participants

A total of 157 student-athletes from one private tertiary institution in Metro Manila participated in the study representing by various courses and different year level. Out of 157, or 87% were male and 70 or *45%* were female whose ages ranged from 16-21 years old with a mean age range of *18*.

The participants were also a combination of rookie, junior, and senior players of various sports such as football, badminton, volleyball, track and field, tennis, softball, fencing, basketball, table tennis, taekwondo, swimming, baseball, judo and chess. *19.10%* are engaged in football, 12 or *7.64%* in badminton, 14 or 8.91% in volleyball, 1 or *0.63%* in track and field, 2 or *1.27%* in tennis, 9 or *5.73%* in softball, 10 or *6.36 %* in fencing, 18 or *11.46 %* in basketball, 10 or *6.36%* in table tennis, 10 or *6.36%* in taekwondo, 10 or *6.36%* in swimming, 3 or *1.91%* in baseball, 20 or *12.73%* in judo and 8 or *5.09%* in chess.

## Procedure

A written permission to conduct a survey on self-efficacy among student-athletes was sought from the director of tertiary level for student athletes covered in this study. It was given to the director with a brief description of the study by the researcher. Notice of permission was given by the director to the different coaches to administer the said test.

## Data Analysis

The survey questionnaire tool for Self-efficacy was constructed and validated first using the Exploratory Factor Analysis (EFA). This procedure was meant to further explore the factors on self-efficacy among student-athletes. The reliability of items in Self-Efficacy Inventory (SEI) was determined using Cronbach's Alpha with an index of .88 which shows good reliability. After the EFA, factor structure was tested using the Confirmatory Factor Analysis (CFA). The CFA approach used was Structural Equations Modeling (SEM) with a maximum likelihood of the variance. Goodness of fit indices was used to test the model for both non-centrality interval estimation and single sample goodness fit indices.

The univariate statistics such as mean, standard deviation, skewness and kurtosis were reported to determine the variability of the measures.

# Results

## Exploratory Factor Analysis

Exploratory Factor Analysis was first used to explore the factor structure of self-efficacy. The Principal Component Analysis procedure extracted four factors (the highest eigenvalues were 11.16 and lowest eigenvalues were 2.94). Using the varimax raw rotation, out of 66 items, 31 were significant, considered acceptable items, and highly loaded in factor marked loading of .50. The total mean score of all 157 respondents along with the

31 items was *M*= 226.74, *SD*= 16.38, skewness= -0.14. The distribution is said to be left skewed, the tail is longer and it has a few low values and kurtosis= 0.039. The consistency of the self-efficacy using Cronbach's alpha was high (.88) which explains the reliability of the items. As a result, a total of 31 items remained and the factors are labeled as: emotional arousal, verbal persuasion, modeling, and performance accomplishments which were based on Bandura's sources of self-efficacy (Weinberg & Gould, 1995).

Table 1
*Accepted Items with their Factor Loadings*

| Item Number | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|---|---|---|---|---|
| 54 | 0.552 | | | |
| 49 | 0.563 | | | |
| 52 | 0.631 | | | |
| 53 | 0.671 | | | |
| 50 | 0.679 | | | |
| 51 | 0.698 | | | |
| 48 | 0.752 | | | |
| 37 | | 0.607 | | |
| 39 | | 0.720 | | |
| 40 | | 0.734 | | |
| 34 | | 0.753 | | |
| 35 | | 0.754 | | |
| 36 | | 0.761 | | |
| 38 | | 0.776 | | |
| 33 | | 0.780 | | |
| 28 | | | 0.500 | |
| 30 | | | 0.511 | |
| 32 | | | 0.518 | |
| 25 | | | 0.531 | |
| 21 | | | 0.533 | |
| 23 | | | 0.551 | |
| 22 | | | 0.574 | |
| 46 | | | 0.568 | |
| 17 | | | 0.579 | |
| 18 | | | 0.619 | |
| 31 | | | 0.634 | |
| 20 | | | 0.646 | |
| 64 | | | | 0.556 |
| 43 | | | | 0.561 |
| 9 | | | | 0.569 |
| 62 | | | | 0.580 |

**Factor 1: Emotional Arousal.** Out of 19 items, seven items were retained and identified as acceptable items; 10 items were removed, while 2 items loaded to the other proposed factors. The factor loading range for this factor from 0.552 to 0.752.

**Factor 2: Verbal Persuasion.** Out of 15 items, 8 items were retained and identified as acceptable items; 5 items were removed whereas 2 items loaded to the other proposed factors. The factor loading range from 0.607 to 0.780.

**Factor 3: Modeling.** Out of 16 items, there were 12 items were retained and identified as an acceptable items and 1 item came from the other proposed factors item. The factor loading range from 0.556 to 0.580.

**Factor 4: Performance Accomplishment.** There were 4 items identified as acceptable items, the factor loading range from 0.556 to 0. 580. Out of the 16 items for this factor, only 1 item identified was acceptable and the 3 items came from the other proposed factors item.

Table 1 shows the accepted items with their factor loadings. The factor means obtained using the sample (n=157) are shown in Table 2. The confidence interval of the means was estimated to determine its accuracy.

Table 2
*Means and Standard Deviation for the Factors of Self-Efficacy among student-athletes (N=157)*

| | N | M | Confidence -95% | Confidence +95% | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Emotional Arousal | 157 | 22.86 | 22.35 | 23.38 | 3.25 | -0.47 | 0.66 |
| Verbal Persuasion | 157 | 28.84 | 28.39 | 29.28 | 2.82 | -1.07 | 1.93 |
| Modeling | 157 | 42.197 | 41.52 | 42.88 | 4.33 | -0.63 | -0.26 |
| Performance Accomplishment | 157 | 11.62 | 11.22 | 12.02 | 2.54 | -0.59 | -0.12 |

The Cronbach's Alpha values of the four factors were .84, .78, .84 and .65 and the overall consistency of the 31 items was .819.

The correlation matrix using Pearson's *r* showed that the factors of self-efficacy have a significant relationship over $p<.05$. Emotional arousal, verbal persuasion, and modeling are significantly correlated to each other however; performance accomplishment did not correlate to any of the three factors.

Table 3
*Correlation Matrix of Self-Efficacy among Student-Athletes*

| Factors | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (1) Emotional Arousal | --- | | | |
| (2) Verbal Persuasion | 0.17** | --- | | |
| (3) Modeling | 0.26** | 0.48** | --- | |
| (4) Performance Accomplishment | -0.09 | -0.01 | -0.05 | --- |

**$p$ <.05

## Confirmatory Factor Analysis

Given the results for the acceptable items in EFA for the four factors of the self-efficacy among student-athletes, it was tested using a confirmatory factor analysis. The goodness of fit based on the RMS standardized residual (RMS=0.070) shows less error having a value closer to .01. The Noncentrality fit indices values shows good fit for self-efficacy four factor (Steiger-Lind RMSEA Index= 0.60, Population Gamma Index= 0.910, Adjusted Population Gamma Index= 0.896).

## Discussion

In the study, four factors of self-efficacy namely: performance accomplishments, modeling, verbal persuasion and emotional arousal were tested among student athletes. The constructed and validated items were first extracted in Exploratory Factor Analysis (EFA) to determine the significant and acceptable items per factors. The EFA factor loading shows high eigenvalues and the number of items was maximized for each factor using the varimax rotation. The total of 31 items retained in EFA were tested using Confirmatory Factor Analysis (CFA). Emotional arousal falls to factor 1, verbal persuasion for factor 2, modeling for factor 3 and performance accomplishment for factor 4. Upon testing the intercorrelations, on the four factors extracted for self-efficacy, three factors showed strong relationships, emotional arousal, verbal persuasion and modeling. However, the performance accomplishment had no significant relationship with the other three factors. This may be in relation with their game exposures wherein some athletes play locally, some internationally, some already engaged in both local and international games and some are not yet engaged in the game though they are already part of the team. Results can be supported by the respondents' profile in relation to their game exposures and experiences. Since the participants of the study were composed of a combination of rookies, and junior and senior players, their mastery experiences may serve as strong evidence as to why it has no significant relationship with the other three factors. Performance accomplishment is part of one's mastery which postulates that if experiences are generally successful, they will raise the level of self-efficacy. However, repeated failures

will result in expectations of lower efficacy (Weinberg & Gould, 1999, 1995 p. 294). In other words, performance accomplishment is in relation with the previous accomplishment of an athlete that needs to be enhanced. Performing a task successfully for athletes may strengthen their sense of self-efficacy. However, failing to effectively deal with those tasks or challenges in their game can undermine and weaken self-efficacy. They should be reminded of personal mastery experiences whether it was successful or not to reinforce the past accomplishment to be able to have a powerful effect on self-efficacy.

Thus, correlated factors such as emotional arousal, verbal persuasion and modeling were identified as significant for student-athletes. Emotional arousal can strengthen self-efficacy by eliciting a situation that could be considered threatening, or that otherwise requires a response from the individual. This may provoke them to respond more strongly than if they did not feel that the situation required a response and can promote self-efficacy and self-esteem if the situation is handled correctly (Bandura, 1994). Athletes' moods, emotional and physical states or reactions, their level of stress or pressure can all influence how they feel about their ability in performing their task. If an athlete feels extremely nervous about his skills and capabilities of what is expected of him, his game may develop a low level of self-efficacy. It is important to note that the emotional arousal of an athlete may make or break a game. However, Bandura also note, "it is not the sheer intensity of emotional and physical reactions that is important but rather how they are perceived and interpreted" (1994, p. 117). By learning how to minimize stress or pressure and lift up mood when facing difficult or challenging tasks, people can improve their sense of self-efficacy.

Verbal persuasion, when coupled with action on the part of the person with low self-esteem or worth, can be a powerful tool in raising self-efficacy (Bandura, 2003, 1977). It is important to note however, that it is more difficult for students to retain self-efficacy bolstered by social persuasion and somewhat easy to cause individuals to doubt themselves and their ability (Bandura, 1994). Verbal persuasion pertains to positive and negative feedbacks that athletes encounter. Getting verbal encouragement from others specifically from their coaches, teammates, family or significant others may help the athletes overcome self-doubt. This may result to them focusing on giving their best effort in their games. This could also influence the athletes to increase their belief that they have the skills and capabilities to succeed in their sport.

Lastly, Ayiku (2005) argued that modeling allows an individual to learn and develop self-efficacy through living an experience vicariously through another (Bandura, 1994). It allows the person to imagine or visualize him or herself in someone's situation. Observing helps a person to like him or herself achieving their goals and be successful in their field. It can also bolster an individual's perspective in accomplishing the same task. Modeling can help others learn essential life lessons and may help those developing self-efficacy in acquiring coping skills to help them complete tasks in the future (Bandura, 1994). One of the important sources of self-efficacy in modeling is when an athlete witnesses other co-athletes successfully completing or performing in a game. According to Bandura, "seeing people similar to oneself succeed by sustained effort raises observers' beliefs that they too possess the capabilities master comparable activities to succeed" (1994, p. 124).

## Implications of the Study

The implications of assessing student-athletes self-efficacy are essential in terms of the four sources used in the study such as performance accomplishments, modeling, verbal persuasion and emotional arousal. It is important to know that these sources play a major role on how student-athletes perceive and how they perform in response to situations, tasks or challenges in their respective sports. Increasing self-efficacy builds positive perceptions of self, which builds an overall self-confidence and creating positive outlook on what is expected of them in performing their sports.

Findings suggest that emotional arousal, verbal persuasion, and modeling have a significant relationship that clearly demonstrates the overall significance of the essence of increased self-efficacy, though performance accomplishment did not correlate with the other three sources. It is important to note for the people who work with the student-athletes in the university such as coaches, managers, sports psychologist and university counselors that athletes should always be reminded of personal mastery experiences whether it was successful or not to reinforce the past accomplishment to be able to have a powerful effect on self-efficacy. Likewise, helping the student-athletes succeed in college and in their respective sports. Results can be a basis in developing programs and modules for the student-athletes.

## Recommendations

It is recommended for future studies to administer the instrument to a larger sample size. Since the study focused only in one private tertiary institution with a combination of rookies, junior and senior players, further study may give more emphasis either for rookie, junior or senior players alone to see its norm and determine its validity in terms of years of engagement in the sports they are into. Further study can also get respondents from the different universities with different sports. With this method, items per factors may increase or may be identified and used as a basis in developing self-efficacy instrument for student-athletes. Whether performance accomplishment did not really correlate or has no significant relationships with the other three factors in the study may also be investigated.

## References

Ayiku, T. Q. (2005). The relationships among college self-efficacy, academic self-efficacy and athletic self-efficacy for African American male football players. Unpublished masters thesis, Graduate School of the University of Maryland, College Park.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior 4* (pp. 71-81). San Diego, Academic Press.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44,* (9), 1175-1184.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37,* 122-147.

Cherry, K. (n.d.) *What is Self-Efficacy?* [on-line site] Retrieved from http://psychology.about.com/od/theoriesofpersonality/a/self_efficacy htm.

Chu R. D., & Tingzon C.J. (2009). The relationship of coaching competency on the athlete's self-efficacy and hope. *The International Journal of Research and Review, 1*, 84-121.

Cox, R. (2002). *Sport psychology: Concepts and applications* (5th ed.). McGraw-Hill Companies, Inc.

Jones, M. V., Mace, R. D., Bray, S. R., MacRae, A. W., & Stockbrida, C. (2002). The impact of motivational imagery on the emotional state and self-efficacy levels of novice climbers. *Journal of Sport Behavior, 21,* 57-73.

Judge, T. A., Jackson, C. L., Shaw, J. C, Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology, 92*(1), 107–127.

Magno, C., & Lajom, J. (2008). Self-regulation, self-efficacy, metacognition, and achievement goals in high school and college adolescents. *Philippine Journal of Psychology, 41,* 1-23.

Mills, K. D., Munroe, K. J., & Hall, C. R. (2000-2001). The relationship between imagery and self-efficacy in competitive athletes. *Imagination, Cognition and Personality, 20* (1), 33-39.

Nicholls, A. R., Polman, R. C. J., Levy, A. R., & Borkoles, E (2010). The mediating role of coping: A cross-sectional analysis of the relationship between coping self-efficacy and coping effectiveness among athletes. *International Journal of Stress Management, 17*(3), 181–192.

Smith, S. A., Kass, S. J., Rotunda, R. J., & Schneider, S. K. (2006). If at first you don't succeed: Effects of failure on general and task-specific self-efficacy and performance. *North American Journal of Psychology*, *8*(1), 171-182.

Vargas-Tonsing, T.M., Warners, A.L., & Feltz, D.L. (2003). The predictability of coaching efficacy on team efficacy and player efficacy in volleyball. *Journal of Sport Behavior*, 24, 396-407.

Weinberg R. S., & Gould, D. (1999). *Foundations of Sport and Exercise Psychology* (2nd ed.). Champaign, IL: Human Kinetics.

## About the Author

**Ms. Maria Cristina M. Firmante** is a registered guidance counselor and a university counselor at the Office of Counseling and Career Services Office of De La Salle University-Manila. She earned her Masters of Education in Guidance and Counseling at the Philippine Normal University-Manila. Currently she is taking her doctorate degree in Counseling Psychology at the De La Salle University-Manila.

Further correspondence regarding this research paper should be addressed to the author, email: maria.cristina.firmante@dlsu.edu.ph and or tina_firmante@yahoo.com.

Educational Measurement and Evaluation Review (2011), Vol. 2, 148-170


http://pemea.club.officelive.com/ EMEReview.aspx

# Beyond assessment: Impact Evaluation of a Community-Based Education Development in Lao PDR

**Benjamina Gonzalez-Flor**
*University of the Philippines - Los Baños*

**Richard DLC Gonzales**
*University of Santo Tomas Graduate School, Philippines*

**Alexander Gonzalez Flor**
*University of the Philippines Open University*

**Abstract** The study assessed the impact of community-based interventions (CBI) that the five-year Second Education Development Project in Lao PDR's poorest districts employed through a quasi-experimental design. Sampled Project village schools were compared with control groups. Control groups were determined using propensity scoring to simulate experimental schools. Treatment or interventions included community-based contracting, community grants, and teacher training. Results of these treatments showed that the CBI package led to positive and improved education outcomes such as increasing enrolment, increased promotion rates, decreased repetition rates, increased gender parity, and higher completion rates. There were significant positive changes in social capital, community development, gender participation, built capacities in village school management, better teaching-learning process and cheaper contracting costs. More significantly, CBI was not only cost-effective but ensured collective ownership as well. The evaluation concludes that effectiveness of one intervention increases in combination with others, making the CBI approach more appropriate.

**Keywords:** *Impact evaluation, community-based interventions, education development, education outcomes, community grants, Lao PDR*

## Introduction

The term *community-based intervention* has a wide range of meaning and seems to pervade not only in developing but developed countries as well (McLeroy, Norton, Kegler, Bardine, & Sumaya, 2003; Pate, Saunders, Ward, Felton, Trost, & Dowda, 2003). It is an activity that is conducted within and by members of a particular community (e.g., grassroots efforts, efforts by a local civic group). It can also be done in partnership with an outside group (e.g., non-government organization, funding agencies, etc). Usually, such interventions occur in various risks mitigation initiatives. Risks in this sense do not only

refer to environmental but all facets of society including education and health (Cole, 2010; Villani & Atkins, 2000). Risks are seen as phenomena that would likely affect one's life drastically or negatively. Hence, any threat to one's survival can be seen as a risk (Pandey & Okazaki, 2005). However, interventions do not seem to be sustainable simply because most of these are top-down in approach (Narayan, 1995).

In a project conducted by the World Bank in 1995, it was stressed that "from time immemorial, communities have organized themselves to take care of collective and individual needs. However, in the last 50 years, so many attempts at getting people to participate and take responsibility for community-based development have failed (Narayan, 1995). Experience provides some clear lessons about what works and what does not work. Prominent among the failures have been attempts to achieve results on a wide scale through the infusion of external management, funds, and technology, controlled from distant places. A fundamental prerequisite of successful participatory programs at the community level is the reversal of control and accountability from central authorities to the community level" (p. 149).

As challenges to community-based organizers and developers escalate, so do the creative new responses that community builders invent (Narayan, 1995). Many of these inventors now recognize that rebuilding low- and moderate-income communities "from the bottom up" requires the mobilization and participation of all of the "assets" at hand. Prominent among these local assets are the local schools. At the same time, local educators are recognizing that successful schools rest on the rock of economically mobile communities.

Thus, there is value to the community-based approach in school development. The need for participation, communication, social cohesion, transparency, accountability, and collective ownership are indeed factors that would determine sustainability of any intervention.

In the case of Lao People's Democratic Republic (PDR), the Ministry of Education (MoE) implemented the Second Education Development Project (EDPII) in 2003 as part of its continuing efforts to achieve MDG number 2 which is the provision of universal access to primary education by the year 2015. The goal of the Project was to increase primary school enrollment and completion in the 19 poorest districts of the six poorest provinces in Lao PDR, namely, Attapeu, Houaphanh, Luang Namtha, Oudomxay, Phongsaly, and Xekong.

EDPII had three components. Component One aimed to increase access and completion of primary education in these areas. It implemented three interrelated community-based interventions (CBI) such as community-based contracting (CBC), community grants for schooling (CG), and teacher in-service upgrading. Component Two focused on the improvement of the quality of primary education through the regular supply of quality textbooks and teachers' guides, and strengthening and increasing the capacity of the MoE in the assessment of student learning outcomes. Component Three addressed capacity building and educational management.

This evaluation study, however, is confined to Component 1 that started in 2005 and ended in 2009. Hence, this study aimed to evaluate the impact of the CBI approach in school development. Specifically, the study aimed to determine the effectiveness of CBI on village beneficiaries as differentiated from comparison groups.

Effectiveness, as used in this study, refers to the comparison of the relative expenditure (costs) and outcomes (effects) associated with two or more alternative courses

of action. In this case, intervention costs were compared with existing costs of school construction in other projects, private companies and communities to determine which one would be the most cost-effective in terms of sustainability over time.

### Results-Based Management Approach in Determining Impacts

Results-based management is a participatory, team-based approach designed to improve program management's effectiveness, efficiency and accountability that focuses on achieving defined results. It is a process of building a culture of measuring outcomes and impacts (goals) as opposed to outputs (deliverables). The evaluation covered three dimensions of project outcomes as evaluation parameters:

**Education outcomes.** Did project interventions improve primary enrolment and completion and reduce repetition in participating villages in relation with comparison villages?

**Targeting precision.** To what extent have project interventions reached the poorest and the disadvantaged?

**Institutional empowerment.** Has the project increased capacity and social capital at the community level in ways that are likely to increase the sustainability and effectiveness of project interventions?

Thus, the study aimed to: determine changes in education outcomes; assess whether project interventions have reached the intended beneficiaries; and find out whether the project resulted to building capacities of stakeholders in village schools and determine whether such capacity is sustainable and effective in managing village schools after the project is over.

## Method

### Evaluation design

The evaluation adopted a non-randomized quasi-experimental design instead of a randomized design covering all six provinces and 19 districts. Baker (2000) argues that a non-randomized quasi-experimental design is the best option when a true-experimental design cannot be done due to several constraints in the field. One shot survey and key informant interviews, focus group discussions and participant observation techniques were employed for triangulation to gather and analyze quantitative and qualitative data. The three stages of evaluation (baseline, midterm and terminal) design were based on the number of Project districts and villages. However, due to synchronization of Project Management Unit (PMU) and project-phasing schedules as well as cost constraints, only two target provinces in the North, and two provinces in the South were included.

## Sampling design

The sampling framework of the study was based on the existing number of Project districts and villages as per EDPII's Phasing-in approach. This included 10 out of the proposed 380 beneficiary villages selected by the PMU and 10 comparison villages covering all six provinces and 19 districts.

The village is the unit of analysis for evaluation. The 10 beneficiary villages and 10 control or comparison villages were identified using *propensity score matching*, in which the comparison group was matched to the treatment group on the basis of a "propensity score" using the Project's eligibility criteria for matching method. The selection of both types of villages included in the sample was based on the criteria identified by the PMU and Project Implementation Design (PID). The comparison villages were selected beneficiary villages. As suggested by Guo, Barth and Gibbons (2004), the study considered criteria that had a total of 100 points. The most important criterion was ethnic representation. Other criteria included location of the school by reaching the remotest which had 20 points each; has not received any external assistance for 15 points; and the rest of the criteria with either 10 or 5 points each being of equal importance.

## Participants

There were four types of respondents in the study: school heads, teachers, village heads, and parents coming from each of the sampled beneficiary and comparison villages. The villages considered were the poorest villages from the poorest districts of 4 provinces in Lao PDR.

There were 20 village heads and 20 school heads interviewed for this study. Respondents also included 24 teachers, at least 1 from each school, 46 parents who were also members of the Village Education Development Committee (VEDC).

## Instruments

This study employed three primary data gathering instruments: an Interview Guide for school heads and village heads; a self-administered questionnaire for teachers, and a focus group discussion guide. Instruments were prepared in English, translated into Lao language, and pretested prior to administration.

Instruments for village heads and school heads included demographic profile, socio-economic characteristics of the village such as population or number of households, gender ratio, ethnic group, sources of livelihood, and activities of the VEDC and level of awareness on EDPII's CBI. The guide for school head also included questions related to school enrollment, dropout rate, repetition rate and children's characteristics. Teacher's SAQ included questions on educational qualifications, in-service training, instructional management and needs for professional development.

## Procedure

Three evaluators gathered the data with the assistance of local researchers who served as interpreters and guide. All interviews and surveys were conducted in Lao and consequently translated into English.

Data gathered from school heads and teachers were validated by District Education Bureau (DEB) chiefs and/or Provincial Education Services (PES) directors. Education outcome statistics were also validated with the Education Management Information System of the Department of Planning and Cooperation of MoE.

Data analyses used descriptive statistics, t-test, one-way ANOVA and bivariate correlation.

## Findings

The presentation of the evaluation results are based on the key indicators or parameters as defined in this study such as education outcomes, targeting precision and institutional empowerment.

### Indicator Number 1: Education Outcomes

Education outcomes include data on average enrolment, completion rate, promotion rate and repetition rate. These indicators are operationally defined as follows: *average enrolment* refers to the actual number of students enrolled in each subject; c*ompletion rate* is defined as the percentage of student completing Grade 5; *promotion rate* indicates the percentage of students moving from one grade to a higher grade level; and *repetition rate* means the percentage of students repeating a particular grade level. It also means non-promotion to a higher grade.

### Enrolment

**Comparison of Average Enrolment Size by Village**. The total baseline enrolment for the 20 selected samples was 816; the midterm was 1,034 while the terminal enrolment was 1,283. The results indicate that the increases of total enrolment of both type of villages from baseline to midterm and midterm to terminal are both significant ($p<.05$), while the increase from baseline to terminal is significant ($p<.001$). Figure 1 shows that the average enrolment size of beneficiary villages is 36.6 in the baseline, 55.2 in the midterm and 70.1 in the terminal. The increase of enrollment among beneficiary villages from baseline to midterm is significant at .05 levels, while the increase from baseline to terminal is significant at .001. Though there was an increase in enrollment from midterm to terminal, the increase, was not significant. For enrollment in the comparison villages, there was no significant increase from baseline to midterm. However, the increase of enrollment from midterm to terminal was significant at .05 while the increase from baseline to terminal was significant at .01.

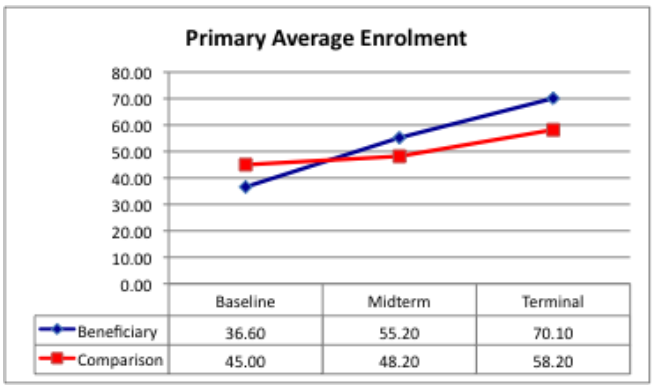



*Figure 1.* Comparison of average enrolment size by village type

**Comparison of Enrolment Size by Gender.** Figure 2 shows the comparison of female enrolment size by village. Initially, the average female enrollment in the comparison villages was higher than the beneficiary in baseline and midterm. While the trend was increasing in both villages, the average enrollment of female in the beneficiary villages increased in the terminal stage. The results also showed that female enrolment in beneficiary villages was more steadily increasing than in comparison villages.

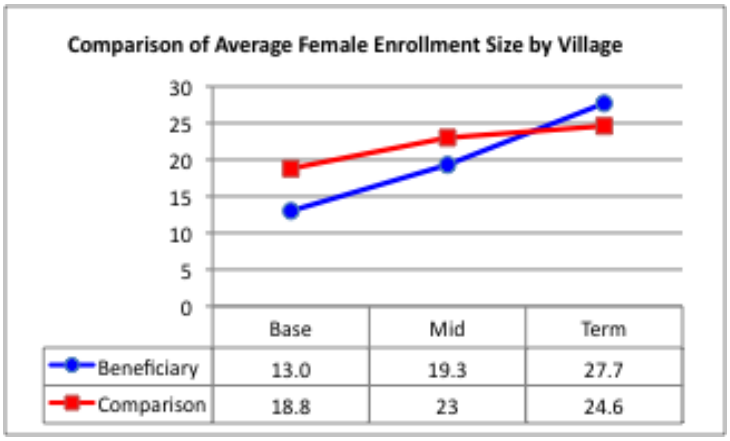



*Figure 2.* Comparison of female enrolment size by village

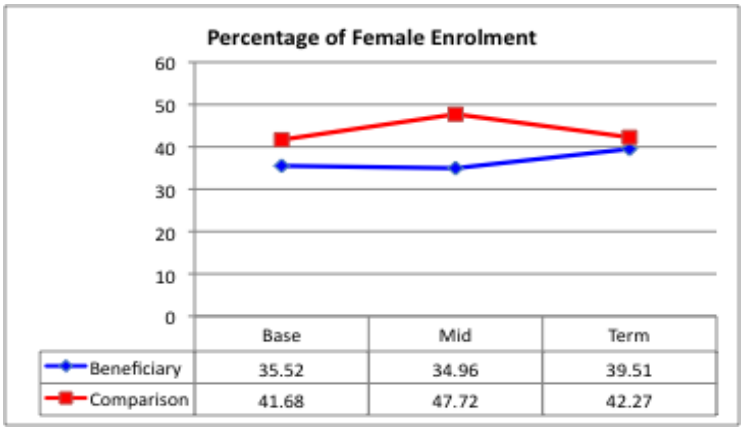



*Figure 3.* Percentage of Female Enrollment

In terms of percentage of female enrollment, in Figure 3 the comparison villages have maintained a significantly higher percentage from baseline to terminal but the difference between beneficiary and comparison at the terminal was no longer different. The results indicate a male-female ratio of 3:2 compared to 3:1 during the baseline, implying further that female participation was increasing in the sample schools.

**Enrollment Growth.** Results showed significant enrollment growth of 57.23 % ($p<.01$) from both beneficiary and comparison villages from baseline to terminal.

The enrollment growth from baseline to terminal among the beneficiary villages was 91.53 % ($p<.001$) and a mere 29.33 % ($p<.05$) in the comparison. Results imply that enrolment in beneficiary villages has grown more tremendously from the time the community-based intervention project was initiated until its fourth year of implementation. However, it was also reflected that the increase from midterm to terminal was not significant. This was because almost all children in the villages have gone back to school during the midterm period when the enrolment growth achieved 50.82 %. The increase can only be attributed to returning older students and new students in Grade 1 (Tables 1 and 2).

Table 1
Enrolment Growth from Baseline to Terminal

| Period | Beneficiary | Comparison | Total |
|---|---|---|---|
| Baseline to Midterm | 50.82%* | 7.11% | 26.71%* |
| Midterm to Terminal | 26.99% | 20.75%* | 24.08%* |
| Baseline to Terminal | 91.53%*** | 29.33%* | 57.23%** |

*$p<.05$, **$p<.01$, *** $p<.001$

Table 2
*Enrolment Growth of Females from Baseline to Terminal*

| Period | Beneficiary | Comparison | Total |
|---|---|---|---|
| Baseline to Midterm | 48.46%** | 22.34%* | 30.01%* |
| Midterm to Terminal | 43.52%** | 6.96% | 23.64% |
| Baseline to Terminal | 113.07%*** | 30.85%* | 64.46%** |

*$p<.05$, **$p<.01$, *** $p<.001$

For total growth enrolment among females was 64.46 % ($p<.01$) from baseline to terminal. The beneficiary villages achieved 113.07 % ($p<.001$) from baseline to terminal while the comparison villages achieved 64.46 % ($p<.01$) only. Results clearly conveyed steady increase of female enrolment in both villages.

**Enrolment of Grades 1 and 2.** Figure 4 shows the average enrolment size of Grades 1 and 2 by village. Overall, Grades 1 and 2 enrolment has been increasing steadily in both villages but the increase was more significant in beneficiary villages from baseline to midterm ($p<.001$). During the baseline, there was no significant difference between the average enrolment size of Grades 1 and 2 for both villages. However, during the midterm, beneficiary villages reported larger enrolment size of Grades 1 and 2 than the comparison villages that posted a growth of 30.16 % ($p<.01$) from baseline to midterm while comparison villages recorded a growth of 5.71 % which was not significant.

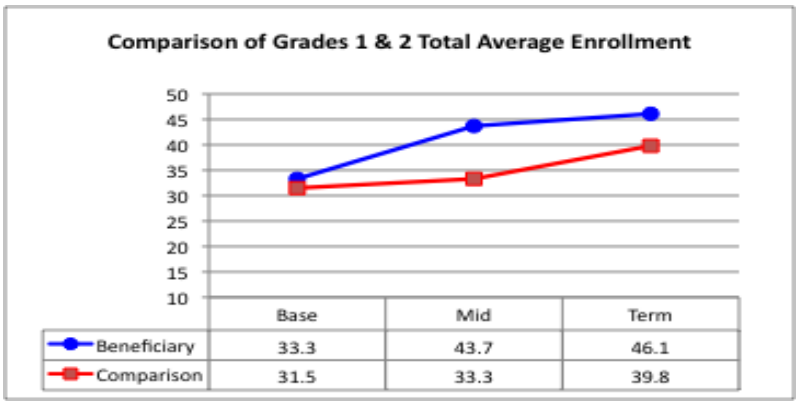



*Figure 4.* Comparison of Grades 1 and 2 total average enrolment size

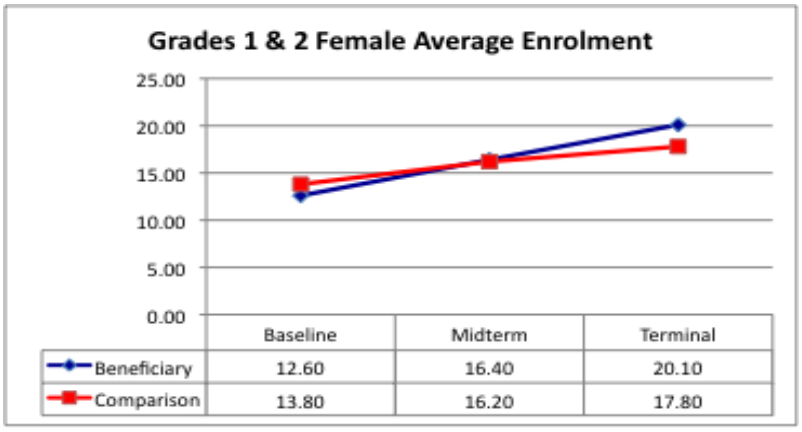



*Figure 5.* Comparison of Grades 1 and 2 female average enrolment size

Female enrolment in Grades 1 and 2 was also analyzed. Figure 5 shows that there was a growth in both villages. However, while beneficiary villages' average enrolment size was lower in the baseline, it increased in the terminal than those in comparison villages (Tables 3 and 4).

Table 3
*Grades 1 and 2 Enrolment Growth from Baseline to Terminal*

| Period | Beneficiary | Comparison | Total |
|---|---|---|---|
| Baseline to Midterm | 31.23%** | 5.71% | 18.83%* |
| Midterm to Terminal | 5.59% | 19.52%* | 11.56* |
| Baseline to Terminal | 38.44%** | 26.35** | 32.56 |

*$p<.05$, **$p<.01$, *** $p<.001$

Table 4
*Female Grades 1 and 2 Enrolment Growth of Females from Baseline to Terminal*

| Period | Beneficiary | Comparison | Total |
|---|---|---|---|
| Baseline to Midterm | 30.16%** | 22.56%** | 23.48%** |
| Midterm to Terminal | 22.56%* | 9.98% | 15.95%* |
| Baseline to Terminal | 59.52%*** | 28.99%* | 43.18%** |

*$p<.05$, **$p<.01$, *** $p<.001$

## Promotion Rate

The promotion rate at terminal stage was 77.16 % compared to 68.98 % during the baseline and 75.53 percent for midterm, indicating a positive growth for both beneficiary and comparison villages. Noticeably, the promotion rates in comparison villages were higher than their counterparts in the beneficiary villages. Among beneficiary villages, the observed promotion rate at terminal was 74.04 % while it was 80.93 % among the comparison villages.

Among females, the promotion rate was higher in comparison villages than in comparison villages from baseline to terminal. However, results also showed that the difference in the promotion rate between beneficiary and comparison was no longer significant at the terminal survey compared to baseline and midterm surveys.

Further analysis of promotion rates from baseline to terminal indicated an increase of 19.78 % ($p<.05$) for beneficiary villages, while only 9.71 % (n.s.) was yielded for comparison villages. For female promotion rate, the increase from baseline to terminal was 28.86 % ($p<.01$) for beneficiary villages while 5.56 % (n.s.) for comparison villages. These results suggest that while the promotion rate of comparison village was higher than the beneficiary villages, the increase of promotion rate from baseline to terminal was more significant for beneficiary villages than comparison villages. Overall, the increase of promotion rate for both villages was 13.50 %.

## Repetition Rate

Another objective of this Project was to reduce repetition rate particularly among the beneficiary villages. The total repetition rate for both villages at the terminal survey was 22.84 % compared to 32.02% at the baseline, posting a decrease of 28.67 % ($p<.01$). Figure 6 shows that repetition rates in beneficiary villages during baseline and terminal surveys were significantly higher than comparison villages. However, during the midterm survey, the repetition rates of both villages were almost the same.

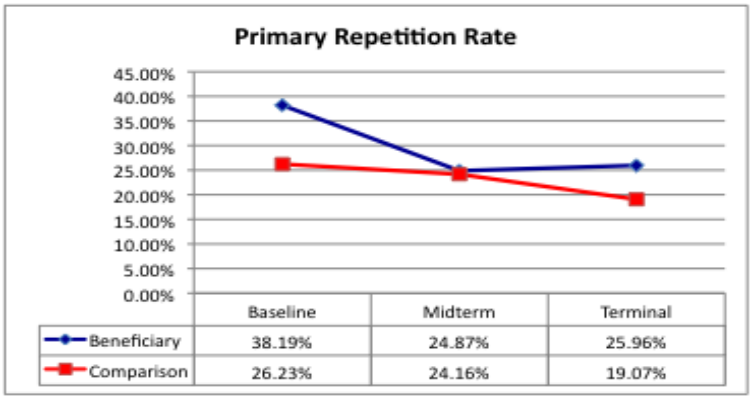



*Figure 6.* Comparison of total repetition rate

The repetition rate of beneficiary villages was 25.96 % at terminal survey compared to 38.19 % at the baseline, posting a significant decrease of 32.02 % ($p<.01$); while the repetition rate of comparison villages was 19.07 % at terminal survey compared to 26.23 %, indicating a significant decrease of 27.29 % ($p<.01$). Though the figure also showed that the repetition rate during terminal survey in beneficiary villages was higher compared to the midterm results, the difference was not at all significant, which still indicates that repetition rate was decreasing from baseline up to terminal survey. A steady decrease of repetition rate, however, was yielded for the comparison villages.

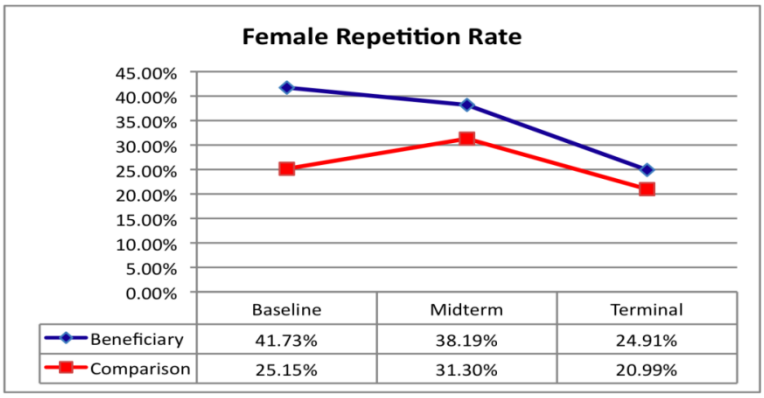



*Figure 6. Comparison of total repetition rate*

For the repetition rate of females, the beneficiary villages have significantly higher repetition rates than comparison villages from baseline to terminal as presented in Figure 6. However, repetition rates of both villages at the terminal survey were no longer significantly different from each other. Both villages have achieved 29.00% ($p<.001$) decrease of repetition from baseline to terminal. The figure also showed that there was steady decrease of repetition rates for the beneficiary village from baseline to terminal, achieving 40.30% decrease ($p<.001$), while the comparison villages achieved 16.54% decrease of repetition rate ($p<.05$). The result clearly implies that female students from both villages have steadily continued their studies and the dropout rate among them was also decreasing very significantly.

## Completion Rate

The analysis for completion was based on the data from 6 complete primary schools, that is, schools that have Grades 1 to 5 classes – three primary schools from beneficiary villages and three primary schools from comparison villages that have complete primary school. Results showed an increase in the completion rate from baseline to midterm. Both villages have achieved a completion rate of 97.76%, and 14.47% ($p<.05$) increased from baseline. The beneficiary villages reached 97.87% completion rate at the terminal, posting an increase of 19.96% ($p<.05$) from baseline; while the comparison villages posted 97.78% completion rate at terminal, an increase of 31.85% ($p<.01$) from baseline.

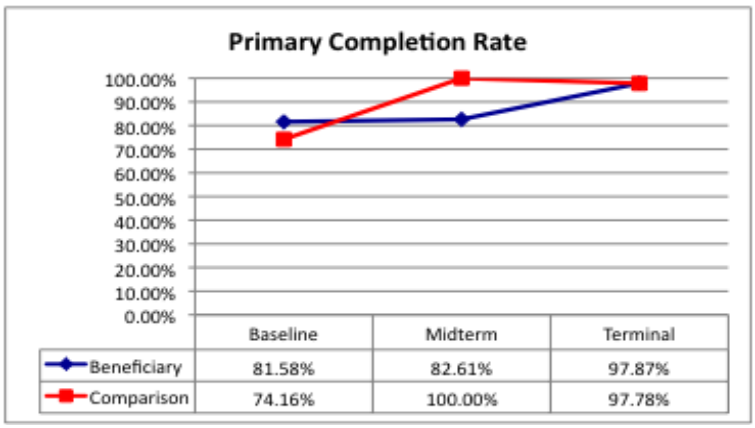



*Figure 7.* Comparison of total primary completion rate

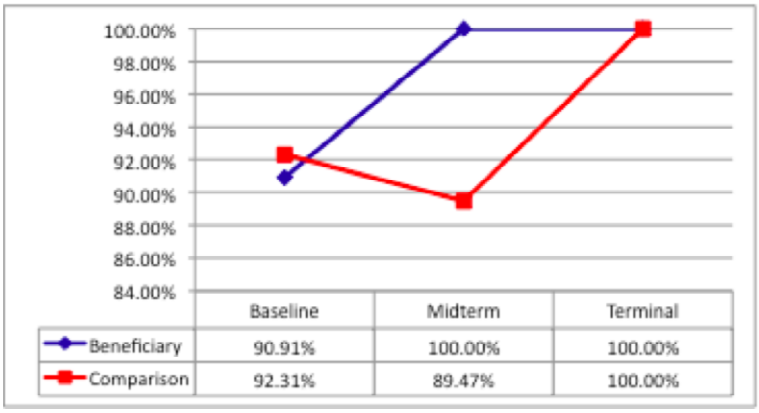



*Figure 8.* Comparison of completion rate for females

The completion rate of females was also determined using the limited number of samples. Noticeably, the completion rate was significantly high, posting 100% for both sample villages at the terminal survey. Results showed that during the baseline survey, the beneficiary villages' completion rate for females was lower than comparison villages'. However, in the midterm survey, the beneficiary villages pulled off a completion rate of 100% and sustained it until terminal survey. The beneficiary villages attained a growth in the female completion rate of 10.00 % from baseline to terminal. For comparison villages, the female completion rate at the terminal survey was 100%, posting an increase of 8.33% from the baseline. Conspicuously, there was a slight decrease of completion at the midterm survey but bounced back to 100% at the terminal survey.

The results imply that students who reached Grade 5 have higher probability to complete primary education. Results also suggest that the government through the Ministry of Education should consider expanding primary education in the village to lower secondary school. Among females, the completion rates imply that they are able to complete primary education given the chance, opportunities and facilities to do so. It can also mean that more females are able to pursue lower secondary education, and eventually upper secondary and even higher education.

## Indicator Number 2: Targeting Precision and Effectiveness

Under Indicator 2, the variables included s*ocio-economic status* operationalized as type of respondents, ethnicity, gender, age, occupation, average monthly income and education expenditures; s*tatus of community-based construction* referring to cost per unit area compared with other school construction models, estimated vs. actual construction costs, and compliance to quality standards; and s*tatus of community grants,* i. e., disbursement and utilization rates of grants.

These differences between the Baseline, Midterm and Terminal Evaluation findings on targeting precision and effectiveness indicators are highlighted and attempts to relate these differences to EDP II's CBI whenever appropriate were done.

## Socio-Economic Status

In this set of indicators, any difference, significant or otherwise, among Baseline, Midterm and Terminal values of the type of respondents, of age and of ethnicity may be considered as functions of non-randomized sampling rather than changes in the socio-economic characteristics of the beneficiaries. Furthermore, differences found in the type, age, and ethnicity of respondents, if any, were close to nil. Hence, this discussion will limit itself to gender distribution, occupation and livelihood, and monthly income. However, it should be reiterated that beneficiary villagers come mostly from the Hmong Mien and Mon Khmer ethnic groups, clearly marginalized sectors in Lao society.

**Gender Distribution.** As with the Baseline and Midterm Studies, gender representation among sectors represented in the midterm heavily favored males. Among the respondents interviewed, 85 % were male. Again, there is a clear indication that women do not generally hold key positions such as village heads. This condition still exists in spite of EDP II's CBI. However, a more sensitive analysis conducted under the educational outcomes and institutional empowerment indicators revealed an improving trend for women representation in the community.

**Parent's Occupation and Livelihood.** All respondents interviewed, including women, were farmers. The Baseline findings gave a wider range of farm related occupations and livelihood among respondents than the Midterm and Terminal findings. The farming sector is one of the marginalized sectors in Lao society. This validates the targeting precision of the Project.

**Monthly Income.** Compared to the Baseline (2005) and Midterm (2007), average monthly household income of Kip 300,000, the estimated average monthly income of both beneficiary and comparison village households was Kip 337,500 or Kip 11250 per day. This amount roughly translates to US$ 1.33 a day. Comparing it to the benchmark, it increased by 33 %. This may be attributable to USD devaluation or inflation in the past four years. Hence, the purchasing power parity (PPP) principle was applied to these comparative figures. However, instead of the usual method of comparing "baskets of goods" that can be bought in 2005 and 2009, respectively, with a predetermined amount or even employing the Big Mac Index, this analysis will use a *rice index* as a gauge.

In 2005, the price of a kilo of rice in rural Laos averaged Kip 4000. The equivalent of the average daily income of beneficiary households of Kip 10,000 was 2.5 kilos of rice. Today, the price of rice averages Kip 4500. The current equivalent of today's average income among the sampled households of Kip 11250 was also 2.5 kilos of rice. Hence, although there was an increase in income in absolute terms, there was no increase in real income based on the current purchasing power of the Kip.

During the Baseline and Midterm Studies, it was likewise argued that US$ 1.00 is what an average rank and file civil servant receives per day. However, nowadays civil servants' salaries have significantly increased due to the GoL's aggressive initiative to upgrade the pay scales of government workers. For instance, primary school teachers are currently receiving Kip 400,000 monthly compared to the 2005 monthly salary of Kip 300-350,000. Additionally, there are other sources of income for the rank and file government employee and his/her members of the family.

Further, when the Millennium Development Goals were drafted in 2000, the international poverty threshold was set at US$ 1 per day, which is more than what the beneficiary families were earning. It was then concluded that EDP II targeted the marginalized and the poorest of the poor in Lao society by international standards.

Nevertheless, given current figures, does the average income of US$ 1.33 per day still validate the targeting precision of the Project? Yes, it does. EDP II continues to benefit the poor segments of Lao society. The current per capita poverty threshold in Lao PDR of Kip 192,000 per month is higher than the average income of beneficiary households translated in per capita terms.

## Community-Based Contracting

**Estimated vs. Actual Construction Costs.** The Baseline findings quoted an ADB estimate of US$ 49.40 per square meter cost of building schools in rural Lao. It was then decided that EDP II's budget estimate of US$60.00 per square meter was adequate given 2004 prevailing prices of construction materials and inflation rates for the next three years. Construction costs per school were, thus, budgeted accordingly. Subsequently, the Midterm Study found that EDP II's actual average cost of school construction under the community-based approach was slightly higher at US$ 62 to 72 per square meter excluding the cost of toilets and water supply. Findings during the Midterm Study prompted the PMU to include toilets and water supply into the cost per unit area estimate, increasing it to an average of US$82 per square meter. Hence, the CBC range based on costs incurred from 2005 to 2007 (when the last EDPII schools were constructed in the South) was US$ 62 to 82 per square meter, *higher* than the 2004 estimate of US$ 60 per square meter.

However, compared to other school construction programs, EDP II had the *least* cost.

**Comparative Costs per Unit Area.** In EDP I, the average cost of building schools in rural areas was US$110 to 130 per square meter. EDP I did not employ the CBC approach then. JICA-GoL schools cost an average of US$ 250 per square meter. However, Japanese contractors were engaged by JICA and Japanese standards of construction and construction materials were adopted. The Second Education Quality Improvement Project, which also engaged contractors, spent US$ 110 to 120 per square

meter. Comparable to EDP II is the Basic Education for Girls Project, which spent US$ 85 to 90 per square meter. However, this figure included school furniture.

The latest school construction program which will be undertaken under the ADB Basic Education Sector Development Project has a cost range of US$190 to 240 per square meter, almost equivalent to the JICA construction cost per unit area. However, BESDP figure represents 2009-10 estimates and has factored in inflation and the weakening US Dollar vis- a-vis the Lao Kip.

**Reasons for Lower Costs.** Based on interviews with key informants at the villages, DEBs and PES as well as with the Deputy Head of ECDM of the MoE Department of Finance, CBC costs were lower due to the following factors:

***Lower overhead costs***. Overhead costs under the community-based approach was close to nil;

***Economies of scale***. District level procurement of construction materials resulted in economies of scale;

***Savings on contractors' fees.*** The reduced role of contractors and the absence of middle men accounted for lower costs;

***Lower labor costs.*** Manpower costs under the community-based approach were substantially lower for obvious reasons. The Project estimate for labor cost under this approach was Kip 50,000 or US$ 5 per square meter (*based on 2005-2007 estimates using 2006 USD to Lao Kip exchange rates)*

***Volunteerism.*** Community-based contracting often used the collective effort approach in building schools, so-called community *sweat equity*." Similar phenomena have been observed in school building projects in Indonesia (*gotong royong*) and the Philippines (*bayanihan*). Community members contributed a counterpart share such as their time, labor and resources in building the facility. These contributions redound to substantive savings in labor and manpower costs within a range of 10 to 90 %. Manpower cost under the community-based approach was estimated at Kip 50,000 per square meter. Hence, savings made from collective effort may be translated to Kip 5,000 to 45,000 per square meter or an average of Kip 25,000 per square meter. Savings on manpower costs were used to build toilets and tap water supply. In this Terminal Evaluation, villagers' *sweat equity* has been estimated at Kip 30,000 – 40,000 per person day.

**Compliance to Standards**. An *Operations Manual* that sets CBC procedures and standards guided community-based contracting under EDP II. It was felt that the use of this manual would ensure efficiency of operations, transparency, as well as quality of construction materials used. Unfortunately, however, CBC may result in non-compliance to professional construction standards. By their very nature, voluntary service inputs and community products (e. g., forest wood vs. kiln dried wood; sand and gravel mix) are not guided by construction industry quality systems.

## Community Grants

**Scheme.** Under EDPII's Community Grants (CG) scheme, beneficiary villages were awarded Kip 30,000,000 annually for three consecutive years. In the course of the

CG implementation, it was found that 10 % of this amount should be devoted for transportation expenses alone and the remaining 90 % for school supplies and equipment.

**Disbursements.** Beneficiary villages in the North (known as Phase 1 villages under the Phasing-in approach) number 267 villages. During the Midterm, community grants have been awarded to all 267 villages beginning 2006. Beneficiary villages in the South (known as Phase 2 villages under the Phasing-in approach) number 52. Community grants for 51 of the 52 villages have been awarded as of the Terminal Study. The grants ceased by the end of School Year 2009-10. All in all, a total of 318 out of 319 beneficiary villages have received community grants: 50 for Luang Namtha, 35 for Phongsaly, 98 for Oudomxay, 84 for Houaphan, 31 for Sekong, and 20 for Attapeu.

**Utilization.** A 99.6 % draw down may be considered a positive indicator for the CGs. However, it likewise revealed the lack of a filtering process that would discriminate effective utilization from poor utilization. This observation was somewhat supported by findings from key informant interviews at the PES, DEBs and villages. There may be a need to incorporate performance-based mechanisms for community grants. A set of evaluation criteria that indicate appropriate and effective utilization should be instituted to determine: continuation or termination; release or withholding of community grant disbursements.

## Education Expenditures

The Baseline and Midterm Studies found that the average annual educational expenditure among beneficiary households was Kip 187,329 and Kip 193,698 per family, respectively. Terminal Evaluation findings showed that education expenditures among beneficiary households significantly increased to Kip 590,000, a 300 % jump. Such was the case even if there was no increase in real income as per computed PPP in the previous chapter.

This 300 % increase in educational expenditure may be attributed to the doubling, tripling, or quadrupling of the number of children within beneficiary households who now attend school. As reflected from educational outcomes, participation rates now approximate100 %. Out of school youth in beneficiary villages are now close to nil. Thus, data on education expenditure support findings on education outcomes.

Beneficiary households tend to invest significantly more in their children's education than comparison households. CGs has encouraged beneficiary families to spend more on education. The Evaluation Team's analysis was limited to community-based interventions and cannot provide findings nationally. However, it can confidently state that at the micro (household) level, there was a significant increase in education expenditure among beneficiary villages.

## Indicator Number 3: Institutional Empowerment

Indicator Number 3 covers the assessment of institutional empowerment and enablement resulting from EDP II's interventions. Conventional wisdom dictates that school management would best be assumed by professionally trained school managers. EDP II differs from other education projects in this respect by handing down such

responsibility to the community through the Village Education Development Committee (VEDC).

The VEDC is composed of the village head, the deputy village head, and representatives of different community organizations such as farmers' groups, youth clubs and women's associations. VEDC is responsible for: the school construction process; planning, implementation and management of community grants; and school management. EDP II believes that communities have the potential to manage schools effectively since they are the primary stakeholders in the education of their children. However, the Project should provide the appropriate interventions initially to empower and enable these communities to manage their schools.

**Gender Participation in VEDCs.** All beneficiary villages have established VEDCs. Women representation in the VEDCs had increased from 13.28 % during the baseline to 26 % in the terminal evaluation. This was due to the required membership composition that included a representative from the Lao Women's Union and the crucial role that they play in community grants disbursement plan. This is a significant move in view of an erstwhile patriarchic structure.

**Frequency of VEDC Meetings.** VEDCs had increased the number of meetings from once a month during the baseline to at least an average of 2.2 times a month in the midterm. This implies that the need to meet more frequently was a critical factor in achieving the goals of EDPII. More consultations, more discussions, more transparency and cooperation seemed to emerge. However, the meetings were back to once a month since mechanisms of implementation had been put in place.

## Social Capital

**Political cohesion/extent of networking.** One of the indicators that would ensure strengthening social capital was political cohesion or establishing networking linkages with concerned organizations or agencies. Stronger political cohesion had been established with government organizations, NGOs, DEBs and other villages. The number of linkages of the beneficiary villages was 71 links during the baseline but significantly decreased during the midterm. The links further decreased to 28 during the terminal evaluation. This change can be attributed to more focus assistance compared to one-shot projects. On the other hand, the comparison groups had 54 links during the baseline evaluation that slightly decreased during the midterm and likewise decreased during the terminal (24). In like manner, assistance given was focused. With the skills gained on school management, proposal writing and financial management, there is a high possibility that villages will submit proposals for funding and thus further strengthen political cohesion. Relationships with DEBs and PES became higher in both types of villages. This implies that education offices are closely monitoring the status of all schools in their jurisdiction.

**Community Participation.** Both beneficiary and comparative villages have worked collectively during the baseline and midterm evaluations. More so, the participation of beneficiary villages was very significant during the terminal evaluation. Participation did not wane either in the comparison villages. This is a good indication that school development is indeed an objective that all villages would like to pursue.

**Community Initiatives for Project Sustainability.** During the baseline evaluation, both beneficiary and comparison villages indicated that they need to support the educational system. This desire was observed to heighten in the terminal evaluation. During the FGDs, they said that information campaign on the importance of education should be sustained. In the absence of the grant, support shall be done on rotation basis. Creation of community-wide income-generating projects such as agriculture and livestock production should be done. Some villages have started community-wide IGPs to realize that support. A village fund through regular collection will be done to be allocated for children's needs and school maintenance. Others have put in place a savings mechanism to ensure that community grants for food especially can be sustained. Since most villages are poor, and parents have to go to the farm and work, food for their children must be made available. In this manner, parents need not worry about their children's lunch. Village heads even required members to send their children to school despite the huge responsibility. They would also strengthen the Parents-Teachers Association. Poor families will be assisted using the village budget. They also plan to prepare a project proposal to be submitted to the district for support. Major repairs, however, is one concern that they may not be able to assume. The arc of potential for sustainability, therefore, is high. It should be considered though those very poor communities would really need some initial assistance until they can put up sustainable income generating projects. The policy of the government in providing block grants would definitely help those communities in the interim.

**Comparative Community Equity.** VEDCs have to establish a mechanism to ensure sustainability of Project implementation. The average annual fixed contributions of beneficiary villages rose generally from Kip 18,941 per family during the baseline to Kip 20,857.14 during the midterm. During the terminal evaluation, fixed contributions varied. Some villages required 10,000 kip per family; others required 168 days of labor; or 38 days of labor depending on the population of the community. Contributions of beneficiary villages in terms of money amounted to 1,470, 000 kip; materials donated amounted to 2,300,000 kip; labor equivalent in the form of sweat equity amounted to 87,920,000 kip; food at 1,650,000; and 400,000 kip for closing ceremonies. The allocation of 3,000,000 kip as community grant indeed helped in easing the burden of parents but not sufficient to support the unintended effects. Younger children now come to school thus increasing the number of children to be fed but at the same time increasing community equity. This implies that beneficiary communities can sustain school management. In the comparison villages, contributions in terms of money amounted to 250,000 kip; materials at 460,000 kip; labor at 14,000,000; food at 2,530,000 kip. Materials donated in terms of kip was difficult to quantify, thus, equity might be higher considering that they will have to build the school yearly for temporary ones; and since there are no permanent schools, they will be expected to contribute for major repairs in the long run. In general, contribution through labor was measured by number of days or square meter of work. In both types of villages, materials (wood) and labor were higher especially with the high cost of materials. Thus, aggregate contributions on the average among the beneficiary villages amounted to 271,093.22 kip per family and 170,150 kip per family for comparison villages. The lower cost in the comparison villages can be attributed to lower cost of temporary construction materials. CBC intervention in this case had increased communities' propensity to assume the responsibility despite the difficulty.

### VEDC Empowerment

**Autonomy in Decision-Making.** During the baseline evaluation, all village heads and community members for both beneficiary and comparison villages decided on day-to-day decision-making and consulted higher authorities for major decisions to be made. With the onset of the Project, more collaboration with government officials and members were required. The process of collaborating increased their ability to think and decide, thus, increasing social capital or the ability to work collectively.

**Leadership Styles.** During the baseline evaluation, 100 % of sample village heads used consultative leadership style and consulted elders for advice. In the midterm evaluation, the preferred leadership style was still consultative but more participatory by soliciting ideas from members especially on deciding on how to spend the CG. VEDCs now closely worked with village constituents for decision-making. During the terminal evaluation, 7 out of 10 used participative decision-making.

**Collective Sense of Project Ownership.** The baseline results found that villages are self-reliant, self-sustained, and self-employed to some degree. School management was one of the activities that they needed to do. During the midterm evaluation, there was an observed strong conviction that villagers have to be in control of project implementation with assistance from DEBs or EDP II officials; more confident that they can solve any problem that may arise during project implementation; and required everyone to understand the importance of education. VEDCs wanted to ensure that the Project is theirs and the responsibility to sustain those lies on them. In the terminal evaluation, not only did they felt that it was their project; they are proud to have built a school as a result of their concerted efforts.

### School Development

In general, there was a significant increase in the number of permanent schools constructed in the beneficiary villages during the midterm evaluation compared to none during the baseline. During the terminal evaluation, all beneficiary schools have permanent structures. On the other hand, the number of semi-permanent structures also increased in the comparison villages. Out of the 10 schools sampled, 8 have permanent structures.

## Discussion

The Evaluation Study concludes that the construction of new school buildings and awarding of community grants in the beneficiary villages have contributed to the improvement of education outcomes. The terminal study results posted a steady significant increase in enrollment and significant growth in the average enrolment size of the beneficiary villages from the baseline survey. Enrolment size in Grades 1 and 2 had increased in both types of villages although the increase was more evident in beneficiary villages. Likewise, female participation has increased significantly for total enrolment and

enrolment in Grades 1 and 2. The gender parity has improved from 3:1 in the baseline to 3:2 in the terminal for total primary enrolment while 1:1 ratio was achieved for Grades 1 and 2 only.

The promotion rate in beneficiary villages showed more significant growth than in comparison villages yielding at least 20 % growth from the baseline data. Further inquiry on this phenomenon from village heads and schools heads indicate that the existence of a more permanent and better school building motivated children to go to school more regularly and encouraged parents to send their children to school. On the other hand, there was a significant decrease of repetition among the beneficiary villages, yielding 32.02 % decrease from the baseline. The female repetition rate of beneficiary villages has also decreased significantly, yielding a decrease of 40.30 %

On targeting precision, the evaluation concludes that the EDPII has targeted the marginalized sectors and the poorest of the poor in Lao society and that community-based construction has resulted in lower school construction costs.

The evaluation can also conclude that Community Grants have been utilized fully with a 99.6 % draw down. However, mechanisms should be instituted to ensure that these are used appropriately and effectively. Although there had been no increase in real income from 2005 to 2009, there has been a significant increase (300 %) in educational expenditure at the micro (household) level due to increased participation rates. Lastly, the evaluation also concludes that beneficiary households tend to invest more on their children's education than comparison households because of the provision of community grants.

On institutional empowerment, the evaluation study concludes that beneficiary communities have been empowered to think of innovative approaches to ensure that every community member gets a decent education towards a better future. Such initiatives are community-wide indicating homogeneity in approach that will ensure sustainability. A mechanism of implementation may need to be put in place like a Village Education Development Fund so that funds are channeled properly. The school now becomes the nucleus of village development. This implies that any development intervention can be coursed through the VEDC. However, VEDC members should have continuing education through non-formal courses or capacity building programs towards developing income generating projects. The type of governance that shifted from consultative to participative type of decision-making did not only lead to social cohesion but assumption of responsibilities as well. The community was enjoined to collectively decide on issues affecting school development initiatives. The most significant change that CBI brought ensured maximum participation not only among village leaders but women, children and the elderly as well. Such collaboration had increased people participation in decision-making to ensure that their aspiration to have quality education as a community is achieved through their own efforts. This sense of self is indicative of strong ownership, pride, and motivation to continue the work that they have started. The study also uncovered a number of positive unintended impacts such as: (i) permanent school buildings now serve as nucleus for basic support services (health, nutrition, etc) in the community and focal point for social development; (ii) permanent school structures serve as a disincentive for nomadic behavior; (iii) there has been a significant decrease in child labor; and (iv) Permanent schools lead to permanent settlements that may result in effective resource stewardship.

On the other hand, negative impact was that the increasing number of primary school graduates cannot proceed to lower secondary education.

In summary, EDPII Community-Based Interventions have functionally achieved their intended outcomes. The evaluation study concludes that these interventions of EDPII were effective.

## References


Asian Development Bank (2008). *Community of practice - Management for development results*. Manila.

Asian Development Bank (2008). *Project implementation manual basic sector education development project*. Manila.

Asian Development Bank (2008a). *Education sector development framework advisory technical assistance for Lao PDR: Sector wide approach in education sector development*. Manila.

Cole, A. G. (2010). School-community partnerships and community-based education: A case study of a novice program. *Perspective on Urban Education, 7*(1), 15-26.

Gonzales, R. DLC. (2010). *Final report: Midterm evaluation of the basic education sector development programme*. Vientiane, Lao PDR: Ministry of Education

Gonzalez-Flor, B., Gonzales, R. DLC., & Lee, R.M. (2005). *Feeling the pulse of the poorest of the poor: A training manual for assessment community-based educational interventions*. Vientiane, Lao PDR: Ministry of Education

Gonzalez-Flor, B., Gonzales, R. DLC., & Flor, A. G. (2005). *Baseline evaluation report: Evaluation of impact of community-based interventions* (Grant No. H0840-LA). Vientiane: Ministry of Education and the World Bank.

Gonzalez-Flor, B., Gonzales, R. DLC., & Flor, A. G. (2007). *Midterm evaluation report: Evaluation of impact of community-based interventions* (Grant No. H0840-LA). Vientiane: Ministry of Education and the World Bank.

Gonzalez-Flor, B., Gonzales, R. DLC., & Flor, A. G. (2009). *Terminal evaluation report: Evaluation of impact of community-based interventions* (Grant No. H0840-LA). Vientiane: Ministry of Education and the World Bank.

Gonzalez-Flor, B. (2009). *Final report: Strengthening education ministry capacity to monitor and evaluate sector programs*. IBRD-IDF Grant No TF090546. Vientiane: Ministry of Education and the World Bank.

Guou, S., Barth, R., & Gibbons (2004). *Introduction to propensity score matching: A new device for program evaluation*. A workshop presented at the Annual Conference of the Society for Social Work Research, New Orleans, January 2004.

Narayan, D. (1995). *Designing community-based development*. Washington, D.C.: The World Bank.

McLeroy, K. R., Norton, B. L., Kegler, M. C., Burdine, J. N. & Sumaya, C. V. (2003). Community-based interventions. *American Journal of Public Health, 93*(4), 529-533.

MfDR *Principles in Action: Sourcebook on Emerging Good Practice*. March 2006. [on-line available] www.capacity.org.

Pandey, B., & Okazaki, K. (2005). *Community-based disaster management: Empowering communities to cope with disaster risks*. Japan: United Center for Regional Development.

Pate, R. R., Saunders, R. P., Ward. D. S., Trost, S. G., & Dowda, M. (2003). Evaluation of a community-based intervention to promote physical activity in youth: Lessons from active winners. *American Journal of Health Promotion, 17*(3), 171-182.

Results-based Management in CIDA (2007). *An introductory guide to the concepts and principles*. Canadian International Development Agency.

Shoshkes, E. (2001). *Smarter planning for schools and communities*. New Jersey: American Planning Association.

UNESCO Asia and Pacific Regional Bureau of Education (2005). *Education for All National Plan of Action 2003-2015.* Ministry of Education, Department of General Education. Bangkok, Thailand.

Villani, C. J., & Atkins, D. (2000). Community-based education. *School Community Journal, 10*(1), 121-126.

## About the Authors

**Dr. Benjamina G. Flor** is a senior faculty member of the UPLB College of Development Communication having served as its College Secretary and Chair of the Department of Development Broadcasting and Telecommunications. Dr. Flor recently completed a year-long assignment in Lao PDR as Team Leader of the World Bank Impact Evaluation of Community-Based Interventions, Team Leader of the World Bank Strengthening the Ministry of Education's Capacity for M&E, and Curriculum Development Specialist of the ADB Basic Education Sector Development Project. Prior to joining UPLB as a fulltime faculty member, Dr Flor has served as Team Leader of the World Bank Community Infrastructure Services Program in Pakistan, Training Activity Manager of the Philippines Australia Short Term Training Facility and Chief of the Monitoring and Evaluation Division of the Commission on Higher Education. Dr. Flor served as Managing Editor of *Medium: A Journal for Philippine Higher Education in Agriculture and Allied Sciences* from 1982 to 1996.

**Dr. Richard DLC Gonzales** is a Professorial Lecturer in the Graduate School of the University of Santo Tomas, Manila, Philippines where he teaches educational and psychological assessment, test and measurement and behavioral statistics. He is currently the President of the Philippine Educational Measurement & Evaluation Association. He is a member of the International Test Commission, Psychological Association of the Philippines and an international affiliate of the American Psychological Association. He also serves as Learning Assessment Specialist for various education projects funded by the Asian Development Bank and the World Bank. He can be contacted by email: r-gonzales@consultant.com

**Dr. Alexander G. Flor is** Professor of Information and Communication Studies at the University of the Philippines - Open University. Formerly UPOU Vice Chancellor for research and development, he was the founding Dean of the Faculty of Information and Communication Studies serving two terms (2004 to 2010). He held the SEARCA-UP Centennial Professorial Chair in 2008-09 and the Metro Manila Professorial Chair in Development Communication in 1995-96. He served as Professor of Strategic

Communication of the University of the Philippines Los Baños College of Development Communication. Dr. Flor completed his PhD in development communication, international relations and policy studies at the University of the Philippines. He was a Fulbright-PAEF post-doctoral fellow at the East West Center Institute of Communication and Culture (Honolulu, 1989). Dr Flor has authored the following books: *eDevelopment and Knowledge Management* (SEAMEO-SEARCA, 2001); *Digital Tools for Process Documentation* (SEAMEO-SEARCA, 2002); *Ethnovideography* (SEAMEO-SEARCA, 2003); *Introduction to Development Communication* (UP Open University, 2003); *Environmental Communication* (UP Open University, 2004); *Development Communication Praxis* (UP Open University, 2007); and *Developing Societies in the Information Age* (UP Open University, 2009). His academic work is featured in Wikipedia.

Educational Measurement and Evaluation Review (2010), Vol. 2, 171-174


http://pemea.club.officelive.com/ EMEReview.aspx

# A Book Review on "Designing Written Assessment for Student Learning"

**Karina M. Agustin**
*De La Salle University, Manila*

**Abstract** The article is a review for the book "Designing Written Assessment for Student Learning" by the Carlo Magno and Jerome Ouano. The book was reviewed in terms of its structure, inclusion of a software in the package, approach and presentation of the book, clarity and organization, and content. The general strength of the book is the suitability of approach for students, contemporary perspective on the topics included, and presentation. Areas that need improvement are on proofreading and proper citations.

**Keywords:** *Self-efficacy, student-athletes*

With the continuous growth of the field of measurement, assessment, and evaluation, several references for this have been published locally one of which is the first edition of the book "Designing Written Assessment for Student Learning" written by Carlo P. Magno and Jerome A. Ouano. This book was copyrighted in 2010 by Phoenix Publishing House, Incorporated, with ISBN 978-971-06-2992-3, and priced at Php 450 (CD included).

The book is designed for students at the tertiary level. It is composed of nine chapters, with subtopics or referred to in the book as "lessons". Each lesson generally contains discussions in paragraph form; however, for certain topics that talks about chronological steps in performing a task, numerical or bulleted format was used. Several tables were also utilized to better organize the data and present the analysis more clearly. Seen at the end of the book is a list of Appendices which shows values and areas relevant for data analysis.

Accordingly, the book is the first in the Philippine publication to include software that facilitates students' learning in analyzing test data. The "Analysis of Test Data Software" is installed in a computer and allows the students to input, process, and view results to be further interpreted and make necessary descriptions and inferences about the test items and the test as a whole. The software may also be used to determine validity and reliability of items or scales. The book highlights the idea of student assessment "for" learning, in contrast with student learning "of" learning. It has been mentioned in the book that the authors are advocates of the idea that the major purpose of assessment is to further guide and help learners in their difficulties and not simply to determine who the better and not so better learners are. The software can perform analysis of reliability that includes test-

retest, split-half, parallel forms, Cronbach's alpha, Kuder Richardson #21, and Interrater reliability. The validity analysis includes criterion-prediction, convergent, divergent, and concurrent validity.

Given that the main theme of the book may be perceived as technical in nature, all of the chapters included exercises and activities to serve as assessments. Questions were also posted in a portion called, "thinking pad" for further analysis of the students on the topic. Every so often, examples for each item or teacher-made test are also provided. This is very helpful in the learning instruction since students at this level still need a basis or a standard example that will be used as a guide for them in the activities. At the end of each lesson and chapter, empirical reports and relevant studies are included for the students to better understand the essence of the discussion. It was also the authors' goal, as stated in the foreword section of the book, to train the students to get accustomed in reading studies and later construct their own hypothesis to be tested using the rigors of assessment.

The book by Magno and Ouano has been written coherently and presented with clarity. It presented the topics in a logical manner, starting off with the overview of assessment, measurement, and evaluation, then towards test development, and ended with the discussion of the status of educational assessment in the Philippines. A similar arrangement or order of topics may also be observed in other similar books. Howeverother booksseldom include empirical reports, relevant studies, and a historical perspective of the educational assessment in the country.

Providing an overview of assessment at the beginning of the book reflects the aforementioned advocacy of the authors to have the audience understand the idea of assessment "for" learning. The process of assessment was given emphasis as well as the nature of measurement and evaluation, although the latter may have been too comprehensively discussed and a bit advanced for the students in the college level. Nevertheless, the empirical study at the end will help the students understand evaluation.

Given several local references published on educational assessments, this book offers contemporary information such as a discussion on the Revised Bloom's Taxonomy. A more comprehensive and updated way of developing learning intents and table of specifications is provided for the students. As I recall being a graduate student in my teacher certificate program, (which was less than 10 years ago) we were trained on lesson plan making and test development using the Old Bloom's Taxonomy. In spite of the advent of the Revised Bloom's Taxonomy, there are still several teachers who rely on textbooks that focus on the Old Bloom's Taxonomy.

In the book, the discussion on the characteristics of an assessment tool (Chapter 3) came before the discussion of test development (Chapter 4). Personally, I would have preferred for the discussion of the characteristics to come after test development. Having taught this course to college students previously, I have found it easier for the students to have a better understanding of the whole idea of determining an assessment tool's reliability, validity, and item analysis after they have mastered the skills of actually constructing their own assessment tool. More value and appreciation may be given to the different characteristics when they gain the knowledge first on how the tool was developed. They realize more the importance of each of the items and of the test as a whole, especially when they constructed those themselves.

As mentioned, the book has greatly emphasized on the characteristics of the assessment tool through determining its validity, reliability, and item analysis. In the process of arriving at such, certain statistical analysis had to be done. The book provides detailed

step-by-step process to determine each characteristic. All the measures were relatively easy to understand. If there is one measure that may need further discussion, it would be the topic on Construct Validity, where factor analysis was used. The discussion on it was quite brief and complex for students, who we should assume may not have gone through any course prior to the assessment course that deals with more than basic statistics. Other than that, everything seemed easy to follow, when asked to do statistical analysis manually. The discussion on the item analysis was particularly the easiest to follow and comprehend.

Generally, the book is very "student-friendly." It provides easy-to-understand discussions given the arrangement of subtopics and the frequent use of bullets and numerical formats. Examples are given every now and then and exercises or activities are also provided – not to mention the answer keys that are included. However, there are some typographical errors in the book such as in some formula (e. g. KR #20 computation), values or data in tables (e. g. table of specifications), and some mistakes in the answer key (e. g. taxonomy exercises) as well. Since I am fortunate enough to know the authors, I was able to approach them for some clarifications on the content of their book. There is still a need for some proofreading to iron out the errors in typing or to even double check the values and text.

In designing written assessment for student learning, constructing test items and developing a teacher-made test as a whole is often the challenge. The book provides general do's and don'ts in writing test items, for both selected response or constructed response types. Besides the guidelines, specific instructions are provided for students to follow. Choice of words for both discussion and examples are comprehensible. Each of the lessons provides references at the end, which students can use as supplements for learning. The content has always been cited and has given due recognition to its reference or original author. I know the content and examples of a lot of educational assessment books are similar, that is why the more proper citing of references is needed. I have encountered a textbook in the past where the content was also the exact content from the internet and was not properly cited.

A unique feature of this book is the inclusion of the CD, which contains the program that facilitates students' learning in analyzing test data. Usually, I am given such a short period to tackle and analyze all the essential topics for an assessment course. More often than not, I find myself unable to use the CD since most of the time we focus on analyzing data manually. Some students do find it more interesting (and of course easier and less time-consuming) to use the program instead. However, it is still personally important to go back to the basics at arriving on an interpretation, especially if it is the students' first time to encounter assessment and statistical analyses. The use of the program to analyze test data can also be used to cross check the initial answers the students obtained from the manual statistics. Moreover, the program will be deemed useful for a more research-driven purpose.

The book's goal is to be able to guide the students in developing their skills on designing written assessments for learning. It can be said that the book has done just that. It has shown to be comprehensive, detailed, and helpful to the students given its "student-friendly" format and logical order of topics. It gave a clear picture on how assessment must be designed through the step-by-step process laid out in the different chapters. It began with the learning intents, characteristics of the assessment tool, writing different types of items, and ways of reporting the results. Moreover, sufficient exercises and relevant empirical studies were presented to give a holistic understanding of each topic.

With these characteristics, I must say the book is very much recommended for students' use and also to teachers and facilitators who teach educational assessments. The book by Magno and Ouano is straightforward and direct and thus, is easy to use. The features of the book's format may be easily adaptable by the students as well. Although the price of (450 php) was not as "student-friendly" as I expected, but I personally I think it is worth the value since it included the software in the package. To resolve such concern of students, an option may be given to consumers alike to choose between a book with the CD or the one without the CD. This then will depend on the need of the consumer, both students and teachers. Other than this, I truly believe that the book, "Designing Written Assessment for Student Learning" by Carlo P. Magno and Jerome A. Ouano, is a substantial and effective reference for educational assessment.

## About the Author

**Ms. Karina Agustin** is presently a faculty of the Counseling and Educational Psychology of the College of Education in De La Salle University, Manila. She teaches courses on Assessment of Learning, Child Development, and Personal Effectiveness. She conducts training for teachers around the Philippines on the conduct of proper assessment. Further correspondence can be addressed to her at karina.agustin@dlsu.edu.ph.