How Subgroup Characteristics Effect Equating Methods' Academic Growth Detection

Ou Zhang
Research and Evaluation Methodology Program, College of Education
University of Florida
119G Norman Hall, Gainesville, FL 32611-7047
E-mail: zhango@ufl.edu, zhangou888@gmail.com

Charles DePascale, PH.D.,
Center for Assessment (NCIEA)
Dover, NH 03821-0351

M. David Miller, PH.D.,
Research and Evaluation Methodology Program, College of Education
University of Florida
119A Norman Hall, Gainesville, FL 32611-7047

Paper Presented at Annual Meeting of the National Council on Measurement in Education (NCME)

annual meeting (2011) at New Orleans, LA

ABSTRACT

The purpose of this study was to, as different levels of growth occur in various size subgroups, investigate the practical consequences of the robustness of the IRT estimation and equating methods in terms of accuracy of proficiency classifications. The test consisted of 47 items including 38 dichotomous items and 9 polytomous items. Grade Response Model (GRM; Samejima, 1969; 1996) and 2PL model were used. IRT true score equating with Stocking-Lord test characteristic curve transformation was applied. Findings indicated that, compared with other factors, the size of the subgroup population (i.e. large subgroup/total population ratio) affected the performance of IRT estimation and equating design most. In addition, regardless of the negative effects from the non-normal characteristics of the total population distribution, true score equating method via Stocking-Lord scale linking approach did play a positive role in recovering the person ability estimates as subgroup growth occurred across years. The study concluded that the decision of choosing the size of subgroup population and different growths in the subgroup population would lead to a very different assessment result of the academic growth for state's large scale assessment.

Key Words: IRT, academic growth, subgroup, equating, under-classification

INTRODUCTION

The No Child Left Behind Act (2001) has increased interest in measuring student Adequate Yearly Progress (AYP) across years (Schwarz, Yen, & Schafer, 2001). Under the No Child Left Behind Act of 2001 (NCLB), each state is required to establish achievement standards and assessment systems for measuring student progress. According to the NCLB Act, federal legislation links funding to standardized test score improvement at Grades 3 through 8. This connection potentially raises the stakes associated with standardized testing throughout the nation (Jodoin, Keller & Swaminathan, 2003). Academic growth and the performance of subgroups are two issues to determine the effectiveness of educational administration and adequate yearly progress (AYP) under the No Child Left Behind Act of 2001.

Academic growth is usually assessed by comparing the performance of students on standardized tests across years (Jodoin, et al, 2003). Overall, the student population for state educational assessments is presumed to be normally distributed. However, differences between subgroups across years within the student population may exist. Subgroup examinees below the proficiency level are often the cause for schools not making AYP. Thus, when a larger proportion of subgroup examinees below the proficiency level are found in the previous academic year, educators usually place extra efforts on lower performing subgroups in the next year to ensure their school to make AYP. Therefore, in this situation, subgroups of students in the second academic year should have more growth than others in the population (Jodoin, et al, 2003). The existence of differential growth by subgroups may impact the shape of the population distribution. As the population distribution changes, suppression or inflation of person ability estimates may occur. So, more growths occurred in the subgroup students may decrease the accuracy of the test item calibration and ability estimation. Larger growth in subgroup may lead to the inaccuracy of the item calibration and ability estimation. It may also influence the accuracy of the population classification, especially

the students from the majority group whose abilities are close to proficiency level cut-offs may be categorized as the "below the proficiency level" students. For instance, students from the majority group whose abilities score are just above the proficiency level, could be categorized as the below proficiency level students, as the large magnitude of subgroup growth occurs.

Very little research has examined how differential subgroup growth impacts population academic growth detection and test form equating. Also, research to examine academic growth detection through equating designs for subgroups of various sizes is limited. Therefore, research about the effect of the subgroup academic growth on the overall population equating and scoring is worthwhile.

For years, the multiple-choice (MC) item format has been the mainstay of standardized testing programs. Recently, mixed-format tests which include both MC and constructed response (CR) items have been earning increasing interests. Because MC and CR items both have their own advantages and limitations, the combination of both item types may allow the concatenation of their strengths while compensating for their weaknesses (Cao, 2008). Therefore, many state assessment systems, including the New England Common Assessment Program (NECAP), have embraced mixed-format tests. Therefore, a mixed-format test including both dichotomous MC items and polytomous CR items, is used in this study.

In the past, Item Response Theory (IRT) has been widely used in the state educational assessment to estimate student performances on standardized tests. Several IRT equating methods are commonly applied to rescale item parameters and equate examinees' performance on the same scale so that academic growth can be detected. Research has shown that the choice of equating methods has large influences on measuring academic growth (e.g. Keller, Keller & Baldwin, 2007). Inappropriate equating methods may result in significant numbers of examinees being placed in

incorrect proficiency categories, and this impact could lead to large consequential results in school's annual AYP assessment .

In summary, this study examined to what extent differential subgroup growths impact population distribution change so that the academic growth detection through commonly applied equating designs (e.g. test-characteristic curve method) and IRT procedures are affected. In addition, this study also investigated the extent to which the equating methods accurately recover population ability across years, for different levels of growth occurring in various size subgroups, such as proportion of the population and amount of growth. Finally, the accuracy of subgroup and the majority group classification across years were appraised as well. Mixed format tests were used in this study. 2PL Model and Samejima's Graded Response Model (1969) were also applied.

THEORETICAL FRAMEWORK

In this section, the theoretical framework of the research is given. The section includes IRT models, equating method, and mixed test format equating weight.

IRT Models

In this study, the Grade Response Model (GRM; Samejima, 1969; 1996) and 2PL model were used. GRM is appropriate to use as item responses are characterized as ordered categorical responses. GRM is an "indirect" IRT model because computing the conditional probability for an examinee responding in a particular category requires a two-step process.

Step One: Operating characteristic curves

$$P_{ijk}^{*}(\theta_i \mid \alpha_j, \beta_{jk}) = \begin{cases} 1 & k = 1 \\ \dfrac{e^{\alpha_j(\theta_i - \beta_{jk})}}{1 + e^{\alpha_j(\theta_i - \beta_{jk})}} & 2 \le k \le K_j \\ 0 & k > K_j \end{cases} \quad (1)$$

Where

α_j is a common slope parameter; the slope parameter is the slope at the point of inflection of the operating characteristic curves.

β_{jk} is a difficulty parameter and its value represents the trait level necessary to respond above threshold j with .50 probability.

The sum of the response probabilities is equal to 0

The higher the slope parameter (α_i), the steeper the operating characteristic curve.

Step Two: Category Response Curves (CRCs)

Category Response Curves represent the probability of an examinee responding in a particular category conditional on trait level.

$$P_{ijk}(\theta) = P_{ijk}^{*}(\theta) - P_{ij(k+1)}^{*}(\theta) \quad (2)$$

The 2PL- Logistic Model can be treated as a special case of GRM.

Scale Transformation and Equating Method

Some equating methods have been examined to detect academic growth, including moment equating methods (e.g. Keller, et al, 2007), test characteristic curve methods (e.g. Hanson & Béguin, 2002), fixed common item parameter methods (FCIP; Paek & Young, 2005), concurrent calibration (e.g. Kim & Kolen, 2006), Tucker linear (von Davier & Wilson, 2008), and equipercentile (e.g.

Doran & Holland, 2000). Results from previous research suggested that the choice of equating methods have essential consequences in measuring academic growth (Jodoin, et al, 2003). Most of research reported that test characteristic curve method and concurrent calibration method appear preferable to the other equating methods in both linking accuracy and robustness (e.g. Kang & Petersen, 2009). Moreover, Béguin, Hanson and Glas (2000) and Béguin and Hanson (2001) reported that under the circumstance of the possible IRT assumption violations, linking using the test characteristic curve methods, especially the Stocking-Lord method, produce more accurate results than concurrent calibration does. Therefore, IRT true score equating via test characteristic curve linking method (i.e. Stocking-Lord) was selected as an equating approach for this study to handle the subgroup growth situations.

Typically three steps are included in the test characteristic curve IRT equating procedures. The first step is item calibration. In this step, appropriate IRT models are used to estimate item parameters of the test (i.e. GRM and 2PL- Logistic Model). The second step is scale transformation to place the estimated parameters from different test forms onto the same scale. In the scale transformation step, Stocking-Lord (Stocking & Lord, 1983) approach as one of characteristic curve methods, is applied for placing IRT parameter estimates from different test forms onto a common scale. The third step is a raw-to-scale score conversion in terms of assigning ability to examinees based on their number-correct scores and corresponding ability estimates from specific score conversion tables.

Data Collection Design

Among data collection designs, nonequivalent anchor test (NEAT) design (Kolen & Brennan, 2004) demonstrates its superiority. In the NEAT design, usually two nonequivalent groups of examinees exist. One group takes form X and the other takes form Y. The anchor test is a set of

common items in both form X and form Y. There are two types of anchors test: external anchor test and internal anchor test. If the anchor section items are counted in the examinee's total scores, the anchor test is referred as an internal anchor. If the anchor section items do not contribute to the examinee's total scores, the anchor test is referred as an external anchor. In the NEAT design, anchor test items are treated as a mini-version of the overall test in terms of content and statistical specifications (Kolen & Brennan, 2004).

Mixed Test Format Scaling Weights

Since a mixed test form was used in the study, applying mixed format test as an anchor in the test equating had its special requirements and challenges. Before implementing equating approaches, the scale transformation must be conducted to place item parameters in the same scale. According to previous research, there are two alternative item weighting approaches for determining an equating line in mixed format test equating (Jodoin, et al, 2003). One item weighting approach is to equally weight dichotomous and polytomous items based on the difficulty parameters from dichotomous items and the location parameter from polytomous items. By using this weighting approach, on the one hand, relatively stable location parameters from the polytomous model are applied; on the other hand, the dichotomous items are over-weighted.

The other item weighting approach is to equally weight the difficulty parameters from the dichotomous items and the threshold parameters from the polytomous items. In this approach, a scoring function is used to associate the scores with the polytomous item categories and dichotomous items. Let W_{jk} refer to the integer score associated with polytomous item category $k-1$. In order to equally scale dichotomous and polytomous items, a response associated with the first category earns a score of 0, a response associated with the second category earns a score of 1, and so forth.

The item response function relates total score on an item to the examinee's ability (e.g. θ). This function is expressed as

$$\tau_j(\theta_i) = \sum_{k=1}^{m_j} W_{jk} p_{ijk}(\theta_i) \quad (3)$$

where $p_{ijk}(\theta_i)$ is the category response function for item j for a polytomous IRT model. For mixed format tests, the test characteristic curve is calculated as

$$\tau_X(\theta_i) = \sum_{j:X} \tau_j(\theta_i) \quad (4)$$

In this study, the second item weighting approach was adopted. Since 2PL- model is a special case of the Graded Response Model (Samejima, 1969), this mixed test equating procedure is also applied to the 2PL- model.

PURPOSE OF THE STUDY

The purpose of this study is three-fold. First, as different levels of growth occur in various size subgroups, to what extent differential academic growths are captured through common equating designs and IRT procedures. The recovery of student ability growth is demonstrated by the subgroup, majority group, and total group mean changes. Second, whether the equating approach recovers the person ability estimates, as suppression, or inflation is found in population ability distribution when subgroup growths exist across years. The person ability estimate recovery of the equating and IRT procedure is demonstrated by the descriptive statistic results under each condition. Third, this study investigates the robustness of the IRT estimation and equating methods in population achievement level classification as the subgroup growths vary in different conditions. This robustness of the IRT estimation and equating methods is indicated by the proportion of the over-classification and under-classification of examinees in the population for each condition.

METHOD

Test Forms

The simulated data sets were generated based on population item parameters from the 2008 New England Common Assessment Program (NECAP) Grade 8 Mathematics Test. The test consists of 48 items including 38 dichotomous items and 10 polytomous items. Among all 10 polytomous items, there is one three-category item in which outlier characteristic is discovered from its parameters. So, this item was removed from the test form for this study. The maximum number of possible points is 64 in this study. Since the mixed format test is considered as an anchor test, the content and statistical representativeness of the anchor test must be taken into account. In previous research, researchers used only dichotomous items in the anchor test for its robustness when the content and statistical representativeness assumptions were violated (Livingston, 1994). However, only using dichotomous items in the anchor test might lead to some serious linking bias. Therefore, a mixed format anchor test should be used to represent corresponding test format and statistical feature of the total test (Kim & Kolen, 2006). Because the purpose of this study is to investigate the equating method's accuracy of the subgroup growth recovery and population distribution changes as subgroup growth occurs across years, the entire test form is applied as the anchor test form for different conditions. Through this design, the effect of discrepancy in content and statistical representativeness between the total test and anchor tests is eliminated.

Classification Cut-scores

In previous research, academic growth was detected by investigating the changes of the percentage of examinees that fell into different classified intervals. As the cut-off score sets as the state proficiency level, the percentage of examinees at or above the proficiency level becomes an important index to report academic growth and AYP (Zhao & Hambleton, 2010). When test cut-

scores are set, the entire examinee population distribution is divided into multiple intervals so that according to his/her latent performance level, every examinee in the population falls into corresponding classified intervals. So the academic growth can be examined in terms of the discrepancies of the percentage of examinees falling in the different intervals across year. On NECAP, there are three cut-scores separating four achievement levels on each NECAP test. Students are classified into one of four achievement levels based on their performance including "Substantially Below Proficient", "Partially Proficient", "Proficient", and "Proficient with Distinction". The cuts are based on the theta-scale, but for any given test there are corresponding raw score cuts. Although the raw score cuts vary a little from year to year, they are usually close to 19, 28, and 48. For the study, scores 19, 28, and 48 are set as classification cut-scores (DePascale, 2006).

Simulation Design

A simulation study was designed using IRT true score equating via the Stocking-Lord linking approach as subgroup growth occurred. Simulations were conducted to investigate the extent to which the equating method accurately recovers group distribution changes when subgroup growths exist across years. The 2PL- model and Graded Response model were adopted for simulation and analysis in the study. The research simulation included 100 iterations with a sample size 20,000.

This research investigated academic growth detection and ability estimate recovery in 3 different conditions:

- Subgroup proportion of the population (subgroup ratio)
- Subgroup mean growth
- Population distribution change

Therefore, three factors were completely crossed: 4 (subgroup ratio)×5(subgroup mean growth) × 2(population distribution change) shown in Table 1:

i. Ratio of subgroup in the student population. Subgroup proportions 0.05, 0.1, 0.25, and 0.5 were applied.
ii. Subgroup ability parameter mean changes including 0 (no growth), 0.25, 0.5, 0.75, and 1.0.
iii. Level of student population distribution change. In this study, 3 different population distribution changes were considered (no changes-normal distributed, mean shift, skewness and kurtosis change).

Mixed Group Ability Normal Distribution Simulation Design and Response Data

A series of samples with 20,000 examinees' ability estimates were randomly simulated for each of the 20 conditions. Population ability parameters for this study were combined from a series of mixed group normal distributions including subgroup and majority group. The null condition subgroup ability parameters were obtained with the mean of -1.3 and standard deviation of 1.1 ($\theta_{subgroup} \sim N(-1.3, 1.1)$). The majority group ability parameters were adjusted according to the subgroup/total population ratio and the subgroup ability parameters under each null condition (i.e. condition 1, 6, 11, 16) so that the population distribution under initial null condition (i.e. condition 1 only) could approximately normally distribute with the mean of zero, and standard deviation of one (i.e. $\theta_j \sim N(0,1)$) within a range from negative four to positive four ($[-4,4]$). The majority group ability parameters remained the same with their corresponding null condition parameters regardless of the subgroup growth occurred across conditions. In the other null conditions (i.e. condition 6, 11, 16), the normally distributed population distribution requirement may not be fully satisfied. This is because the increases of subgroup/total population ratio may result in the complex issue for the

mixed group normal distribution simulation. However the mean of zero for the null condition population distribution was ensured as the minimum requirement of the simulation for this study. The subgroup/majority group population density distribution examples for the first 5 conditions were displayed in Figure 1. Five subgroup ability distributions were included in this study (i.e., $N(-1.3, 1.1)$, $N(-1.05, 1.1)$, $N(-0.8, 1.1)$, $N(-0.55, 1.1)$, and $N(-0.3, 1.1)$).

For the purpose of comparison, the mean-shift conditions for the population ability distribution were also applied accordingly. The mean-shift condition kept the population ability normally distributed, but shifted the mean of the distribution across the conditions, while the standard deviation as one was retained. For the simplicity purpose, all simulated population parameters were rounded to three decimal places. Table 2 shows the descriptive statistics of the population ability for simulation per each condition. The population ability density distributions of all the conditions are shown in Figure 2.

Shown in the Table 2, the means of the population ability for condition 1, 6, 11, and 16 (i.e. 'no subgroup growth' conditions) were around zero. The standard deviations of these 'no subgroup growth' conditions were around one except condition 16. This is because in data simulation, the subgroup mean and standard deviation values remained at -1.1 and 1.3 respectively, as the subgroup/population ratio increased from 0.05 to 0.5, two groups of populations (i.e. subgroup and majority group) spread out the total population distribution in condition 16 to keep the total population mean as zero. Consequently, bimodality occurred in the total population distribution. For this reason, the total population standard deviation increased from 1 to 1.65 in condition 16.

The population ability distribution for most of the conditions showed negative skewed and leptokurtic characteristics, as the subgroup/population ratio and subgroup growth were small. When the subgroup/population ratio increased to its extreme, the population ability distribution shape

turned to platykurtic. The descriptive statistics and population distribution for Mean-shift conditions are shown in Table 3 and Figure 3 as below.

Calibration of the Data and Estimation of Abilities

In the study, item parameter calibration was analyzed using Marginal Maximum Likelihood (MML) methods with the 'ltm' R package (Rizopoulos, 2006). The estimation of the simulees' abilities was conducted by Expected a Posteriori Estimation (EAP; Bock & Mislevy, 1982).

Procedure

First, a series of data simulations were conducted. Second, item calibration was operated via 'grm' function from 'ltm' package (Rizopoulos, 2009) on these simulated item responses. Third, a set of IRT scale transformation coefficients was obtained by using the Stocking-Lord test characteristic curve approach via 'plink' package (Weeks, 2010). Fourth, IRT true score equating was applied to obtain the true score and ability estimate. Fifth, the mean and standard deviation assigned ability estimates per conditions were calculated. The percentages of examinees falling into the different classification intervals were also calculated. The mean and standard deviation expected growth were compared to evaluate how the equating method captures the differential subgroup growths. The discrepancies between classification percentages of examinee ability obtained from a converted raw-to-scale table after equating and classification percentages of examinee ability obtained from the original ability distribution are obtained to evaluate how accurately the equating method recovered group distribution changes. Finally, the accuracy of category classification was assessed for both subgroup and majority group ability estimates compared with their corresponding population ability values.

Evaluation Criteria

The estimates from the 'no subgroup growth' conditions were transformed to the metric of the true parameter values. In this study, the Stocking-Lord linking transformation method was

applied and the entire test form was used as the anchor test. Then, each subsequent condition is equated backwards to the corresponding 'no subgroup growth' condition (i.e. same subgroup/population ratio) using the same scale linking transformation method.

Measure of Growth

The amount of subgroup growth recovered in each condition was indicated by a mean difference between the subgroup examinee's ability estimated after equating and the original true ability for different conditions. The other measure of growth was assessed by the classification of examinees into performance categories. A decrease in the number of examinees classified as "Substantially below proficient" (e.g. below cut-score 19) could be interpreted as a measure of academic growth. Therefore, the discrepancy of ability estimate classification for subgroup population between the ability estimates after equating and original ability was investigated.

Under-Classification and Over-Classification

The consistency of classification for subgroup and majority group ability parameters with regard to different subgroup ability growth was used as the criterion to evaluate the recovery capability of the equating approach and IRT calibration on population ability. The percentages of over-classification and under-classification for subgroup and majority group examinees were summed as the classification inconsistency coefficient to indicate the capability of the equating approach detecting academic growth. Over-classification means the examinee's ability is over classified into a higher category. For instance, an examinee's true ability is just below proficiency level and should be classified in the below proficiency interval. However, because of inaccuracy of estimation or equating procedure, this examinee's ability is estimated above proficiency level and classified into the above proficiency level interval. Under-classification is the opposite scenario of over-classification. In the realistic situation, under-classification negatively impacts on the results of the state education performance so that AYP evaluation of the states is biased.

RESULTS

Descriptive Statistics and Density Distribution

As previous noted, it is necessary to look at the descriptive statistics and density distribution of the ability estimates for all the conditions. These results are shown in Table 4 and Figure 4. The results in Table 4 indicated that although the population ability mean for all condition varied, the means of ability estimates remained around zero. Meanwhile, the magnitudes of the ability estimates standard deviation shrank (i.e < 1.0). The shrinkage of the ability estimates standard deviation occurs due to the EAP property. The ability estimates distributions for all the conditions turned from negative skewed to positive skewed. The ability estimate distributions for all the conditions held the leptokurtic characteristics, except the extreme subgroup/population ratio conditions (i.e. condition 16-20).

A possible cause of the zero value of ability estimate means is that MML was applied in item calibration. In the MML method, the quadrature nodes and corresponding weights are assigned from a normally distributed prior distribution. In this study, updated posterior ability distribution was not considered for two reasons. First, the purpose of this study included how the differential subgroup growth affects the extent of item calibration and equating approach in academic growth detection. As the false normal distribution assumption holds by the default item calibration procedure, the accuracy of academic growth detection may decline. Second, in the realistic situation, the extreme subgroup/population ratio (0.5:1) may not occur.

Similar to the population ability distribution, as the subgroup/population ratio increased to its extreme, the ability estimate distribution shape turned to platykurtic. The descriptive statistics

and density distribution of ability estimates for mean-shift conditions are shown in Table 5 and Figure 4 as below.

Ability Estimates Mean difference

The magnitude of the ability estimate mean difference is one of the important indices indicating the amount of academic growth recovery capability of equating approach. Table 6 lists the ability estimate mean difference of subgroups and majority groups for all 20 conditions. The difference between expected mean growth and observed mean growth indicated the robustness of IRT estimation procedure with prior normal distribution and the equating method on the academic growth detection.

As we can observe from the table, the results of the observed mean growth for subgroup showed that the equating method and IRT estimation procedure with prior normal distribution captured half of the subgroup academic growth when the subgroup/population ratio was small (i.e.0.05:1). As the subgroup/population ratio increased, the accuracy of subgroup academic growth recovery by the equating method and the IRT estimation procedure declined. When the subgroup/population ratio reached to its extreme condition, the accuracy of subgroup academic growth detection by the equating method and the IRT estimation procedure declined to its minimum but the trend of academic growth continued.

In the meantime, the majority group ability estimates mean difference held within a small magnitude level across different subgroup/population ratios. This indicated the robustness of ability estimation and equating methods on the no academic growth majority groups. When the subgroup/population ratio held constant, the majority group ability estimates mean difference decreased but the absolute value of its estimate increased, as the subgroup growth increased. This indicated that the majority group ability estimates are negatively influenced by the magnitude of

subgroup growth. As the subgroup/population ratio increased, the negative impact of the majority group ability estimates by the magnitude of subgroup growth increased.

Under-Classification and Over-Classification

The classification to put examinees into different categories based on their ability performance is an important index to assess how the differential subgroup growths affect the IRT procedure and the equating method's academic growth detection. Table 8 shows an example of the classification contingency table for conditions. The results shown in Table 8, display the classification of total population for null condition 1. The results in the diagonal cells indicated the number of correct classified examinees among total 20,000 examinees. The number in the upper right triangle cells revealed the number of over-classified examinees. As we can observe in the table, the under-classification and over-classification occurred in the null condition. The possible reason of these under-classification and over-classification occurrences might be the variation of the mixed group normal distribution simulation. The number in the lower left triangle cells indicated the number of under-classified examinees.

By summing all the number of over-classified examinees in the upper right triangle cells and divided by the corresponding sample size, the proportion of over-classification for each condition is calculated. Likewise, the proportion of under-classification for each condition is calculated, by summing all the number of under-classified examinees in the off diagonal lower left triangle cells and divided by the corresponding sample size. Table 9, Table 10, and Table 11 provide all conditions' over-/under- classification proportions results for the total population, subgroup, and majority group after equating.

Table 9 displays the under-/over- classification proportion results for the entire population. As we can observe from the table, several trends appeared to hold. The proportion of over-classification for all conditions was well-controlled around 5 percent (0.05). As the subgroup

growth increased, the over-classification proportion decreased accordingly. Meanwhile, as the subgroup/total population ratio increased, the over-classification proportion decreased as well. However, compared with over-classification, under-classification drew more problems. As the subgroup/total population ratio remains low (≤ 0.1), the proportion of under-classification for conditions varied from 5 percent (0.05) to 10 percent (0.1). As the subgroup growth increased, the under-classification proportion increased accordingly. Meanwhile, as the subgroup/total population ratio increased, the under-classification proportion increased as well. Specifically, as the subgroup/total population ratio reached over 0.25, the under-classification proportion increased drastically. The maximum proportion of the under-classification reached over fifty percent (0.5276) as the subgroup growth and the subgroup/total population ratio reached to their maximum (subgroup growth 1.0; subgroup/total population ratio 0.5:1).

Table 10 displays the under-/over- classification proportion results for the subgroup population. Similar to the trends held in the total population, the proportion of over-classification for all conditions was well-controlled around 5 percent (0.05). As the subgroup growth increased, the over-classification proportion decreased. As the subgroup/total population ratio increased, the over-classification proportion decreased as well. Similar to the trends held in the total population, as the subgroup/total population ratio kept low (≤ 0.1), the proportion of under-classification for conditions varied from 1 percent (0.01) to 10 percent (0.1). As the subgroup growth increased, the under-classification proportion increased. Meanwhile, as the subgroup/total population ratio increased, the under-classification proportion increased as well. The magnitude of the under-classification proportion increase for subgroup was not as large as its corresponding increase for the total population. Specifically, as the subgroup/total population ratio reached over 0.50, the under-classification proportion increased drastically. The maximum proportion of the under-classification reached around forty-seven percent (0.4780) as the subgroup growth and the subgroup/total

population ratio reach to their maximum (subgroup growth 1.0; subgroup/total population ratio 0.5:1).

Table 11 displays the under-/over- classification proportion results for the majority group population. Similar to how the trends held in the total population, the proportion of over-classification for all conditions was well-controlled around 5 percent (0.05). As the subgroup growth increased, the over-classification proportion decreased. As the subgroup/total population ratio increased, the over-classification proportion decreased too. Slightly larger trends held for the majority group compared with total population trend, the proportion of under-classification for conditions varied approximately from 6 percent (0.0582) to 14 percent (0.14), when the subgroup/total population ratio kept low (≤ 0.1). As the subgroup growth increased, the under-classification proportion increased accordingly. Meanwhile, as the subgroup/total population ratio increased, the under-classification proportion increased as well. The magnitude of the under-classification proportion increase for the majority population was larger than its corresponding increase for the total population and subgroup population. Specifically, as the subgroup/total population ratio reached over 0.25, the under-classification proportion increased drastically. The maximum proportion of the under-classification reached over fifty-seven percent (0.5772) as the subgroup growth and the subgroup/total population ratio reached to their maximum (subgroup growth 1.0; subgroup/total population ratio 0.5:1).

Table 12, Table 13, and Table 14 provide all the mean-shift conditions' over-/under-classification proportions results for total population, subgroup, and majority group after equating, respectively. Table 12 displays the under-/over- classification proportion results for the entire population. The proportion of over-classification for all conditions was well-controlled around 5 percent (0.05). As the subgroup growth increased, the over-classification proportion decreased drastically to zero. Meanwhile, as the subgroup/total population ratio increased, the over-

classification proportion for each null condition varied, but decreased drastically when the subgroup growth occurred. However, compared with over-classification, under-classification drew more problems. As the subgroup/total population ratio remained low (≤ 0.1), the proportion of under-classification for conditions varied from 7 percent (0.0674) to eighty-five percent (0.8522). As the subgroup growth increased, the under-classification proportion increased drastically. Meanwhile, as the subgroup/total population ratio increased, the under-classification proportion increased drastically. The maximum proportion of the under-classification reached over eighty-five percent (0.8522) as the subgroup growth and the subgroup/total population ratio at subgroup growth 1.0 and subgroup/total population ratio 0.05:1.

Table 13 displays the under-/over- classification proportion results for the mean-shift condition subgroup population. Similar as the trends held in the total population, the proportion of over-classification for all conditions was well-controlled around 5 percent (0.05). As the subgroup growth increased, the over-classification proportion decreased accordingly. As the subgroup/total population ratio increased, the over-classification proportion decreased too. As the subgroup growth increased, the under-classification proportion increased accordingly. The magnitude of under-estimation proportion for all null conditions (no subgroup growth) was comparatively low (around 0.05). However, the magnitude of under-estimation proportion increased drastically when the subgroup growth occurred. Meanwhile, as the subgroup/total population ratio increased, the under-classification proportion increased as well. The magnitude of the under-classification proportion increase for subgroup was not as large as its corresponding increases for the total population, but it still kept within a fairly high level. Specifically, as the subgroup/total population ratio reached over 0.50, the under-classification proportion increased drastically. The maximum proportion of the under-classification reaches around fifty-five percent (0.5502) as the subgroup growth and the

subgroup/total population ratio reached to their maximum (subgroup growth 1.0; subgroup/total population ratio 0.5:1).

Table 14 displays the under-/over- classification proportion results for the mean-shift condition majority group population. Similar to how the trends held in the total population, the proportion of over-classification for all conditions was well-controlled around 5 percent (0.05). As the subgroup growth increased, the over-classification proportion decreased drastically. As the subgroup/total population ratio increased, the over-classification proportion decreased too. Slightly larger trends held for the majority group compared with total population trend, the proportion of under-classification for conditions varied approximately from seven percent (0.0704) to approximately seventy-eight percent (0.7787), when the subgroup/total population ratio kept low (≤ 0.1). As the subgroup growth increased, the under-classification proportion increases. Meanwhile, as the subgroup/total population ratio increases, the under-classification proportion increased as well. The magnitude of the under-classification proportion increase for majority group was larger than its corresponding increases for the total population and subgroup population. The maximum proportion of the under-classification reached over eighty-four percent (0.8406) as the subgroup growth and the subgroup/total population ratio at subgroup growth 1.0 and subgroup/total population ratio 0.25:1.

DICUSSION AND CONCLUSION

One purpose of this study was to evaluate the effect of subgroup growth on the performance of common equating designs and IRT procedures, as different levels of growth occur in various size subgroups. The other important purpose of this study was to assess the robustness of the IRT estimation and equating methods in population classification. The results suggested that the size of the subgroup population (i.e. large subgroup/total population ratio) affects the performance

of IRT estimation and equating design most, compared with other factors. It was found that while there was no subgroup growth, commonly used equating approach and IRT estimation have a very low performance when the subgroup/total population ratio was large. Meanwhile, the results of mean-shift condition indicated a worse scenario than the conditions in which only subgroup growth existed. This phenomenon indicated that the non-normal characteristics of the total population distribution negatively affected the performance of defaulted IRT estimation (i.e. normally distributed population distribution assumption is hold) even before the equating approach was applied. As the shape of the total population distribution kept approximately normal, the equating approach was able to detect certain amount of the subgroup academic growth. But when the shape of the total population distribution became non-normal, the equating approach had a very poor performance on the academic growth detection.

Since misclassification directly relates to the AYP evaluation, this measure is of utmost importance. Specifically, because the under-classification negatively impacts on accountability results, it must be assessed carefully. There was a pattern of results of over-classification and under-classification across conditions. On a related note, the size of the subgroup population influenced the over-classification and under-classification most. As the subgroup/total population ratio increased, the over-classification proportion decreased and the under-classification proportion increased dramatically for all conditions under both subgroup growth and mean shift circumstances. Specifically, as the subgroup/total population ratio reached over 0.25, the under-classification proportion inclined drastically. The maximum proportion of the under-classification usually reached its maximum as the subgroup growth and the subgroup/total population ratio pulled to their extremes.

On the one hand, there was high proportion of under-classification when subgroup/total population ratio was large, regardless of subgroup growth changed. On the other hand, if the

subgroup/total population ratio kept at comparatively low level (i.e. ≤ 0.1), over-classification proportion declined and the under-classification proportion inclined, as the subgroup growth increased.

Regardless of the negative effects from the non-normal characteristics of the total population distribution, true score equating method via Stocking-Lord scale linking approach did play a positive role in recovering the person ability estimates as subgroup growth occurred across years. As we can observe in a test characteristic curve (TCC) example from Figure 9, the equating approach pulled the growth-existing condition TCC (i.e. red dash growth curve line) back (i.e. green dot curve line) to null condition TCC (i.e. blue curve line) as subgroup growth occurred. In this study, because of the huge negative impact from the non-normal characteristics of the total population distribution on the IRT estimation, the positive effect from the true score equating method in person ability estimates recovery was lessened.

In this study, a limitation exists regarding the method of multi-group mixed normal distribution simulation. Different ways of simulating multi-group mixed normal distribution play a key role in this study. Clearly, set by design, the current research results indicated that data simulation in this study was not fully satisfactory to meet the research design as we initially planned. But multi-group mixed normal distribution simulation always exists as the main issue in the simulation research. Thus, future studies should consider assigning quadrature nodes and corresponding weights for the population distribution according to the particular research design.

The other limitation of IRT estimation also relates to the non-normal population distribution characteristics. In this study, the prior ability distribution was set as default normal distribution for MML in IRT estimation. The default prior normal distribution was set to match the circumstance as the usual procedure in state's large scale assessment. By setting prior distribution as normal, the

IRT estimation applied in this study was hugely negative influenced by the non-normality of simulated population distribution. Thus, future studies should consider posterior ability distribution updates (Paek & Young, 2005) or nonparametric IRT approach (Sijtsma, 2002) in the IRT estimation step.

In a nutshell, the results of this study mainly met its purpose to evaluate the extent of differential academic growth captured through common equating designs and IRT procedures. In this study, the non-normality and the size of subgroup population issues merged. Therefore, the decision to choose the size of subgroup population would lead to a very different assessment of academic growth for the state's large scale assessment. Inappropriate subgroup population sample size selection may raise questions as to the appropriateness of the results of equating method academic growth detection. Therefore, it is important to consider the size of subgroup population and the distribution of population in the academic growth detection analysis.

References:

Béguin, A. A.,&Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Béguin, A. A., Hanson, B. A.,&Glas, C. A.W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Bock, R. D.,& Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Cao, Y. (2008) . Mixed-format test equating: effects of testing dimensionality and common-item sets. *Dissertation submitted to the Graduate School of the University of Maryland, College Park.*

DePascale, C. (2006). Establishing Academic Achievement Standards for the New England Common Assessment Program. *Technical Paper produced by the "New England Common Assessment Program".*

Dorans, N.J. & Holland, P.W. (2000). Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case. *Journal of Educational Measurement, 37 (4),* 281-306

Harris, D.& Kolen, M.J. (1986).Effect of Examinee Group on Equating Relationships. *Applied Psychological Measurement* (10)35

Hanson, B., & Béguin, A. A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement* (26) 3

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71,* 229–250.

Kang.T. & Petersen., N.S. (2009). Linking Item Parameters to a Base Scale. *ACT. Tech Report.*

Keller, R.R., Keller, L.A. & Baldwin, S. (2007).The Effect of Changing Equating Methods on Monitoring Growth in Mixed-format Tests. *Paper presented at 2007 NCME annual conference.*

Kim, S. & Lee, W-C (2004). IRT Scale Linking Methods for Mixed-Format Tests. *ACT Research Report Series 2004-2005*

Kim, S & Kolen, M. J.(2006).Robustness to Format Effects of IRT Linking Methods for Mixed-Format Tests., *Applied Measurement in Education, 19(4),* 357 – 381

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Livingston, S. A. (1994). *Equating constructed-response tests through a multiple-choice anchor: A small-scale empirical study*. Unpublished statistical report.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Paek, I. & Young, M. J.(2005). Investigation of Student Growth Recovery in a Fixed-Item Linking Procedure with a Fixed-Person Prior Distribution for Mixed-Format Test Data., *Applied Measurement in Education, 18(2),* 199 — 215

Peterson, N. S. (2008). A Discussion of Population Invariance of Equating. *Applied Psychological Measurement, 32*, 98-101.

Rizopoulos, D. (2006) ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17(5),* 1–25. URL http://www.jstatsoft.org/v17/i05/

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No.17.

Samejima, F. (1996). The graded response model. In W.J.van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory*. New York: Springer.

Schwarz, R.,Yen, W.,Schafer, W. (2001). The Challenge and Attainability of Goals for Adequate Yearly Progress. *Educational Measurement: Issues and Practice. 20(4),* 26-33.

Sijtsma, K. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, Calif.: SAGE Publications, c2002

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

von Davier, A. & Wilson, C., Investigating the Population Sensitivity Assumption of Item Response Theory True-Score Equating Across Two Subgroups of Examinees and Two Test Formats, *Applied Psychological Measurement* 2008; 32; 11

Weeks, J. P. (2010) plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35(12),* 1–33. URL http://www.jstatsoft.org/v35/i12/

Zhao, Y., Hambleton, R. (2010). Practical consequences of model misfit in assessing academic growth. *Paper presented at the annual meeting of the American Educational Research Association (AERA)*, Denver, CO.

APPENDIX A: TABLES

Table 1. Condition Table (Subgroup Mean Changes and Subgroup Proportion Changes)

Subgroup mean and SD Change						
		Mean Growth Level				
		0	**0.25**	**0.5**	**0.75**	**1**
Subgroup: Population Ratio	**0.05**	1	2	3	4	5
	0.1	6	7	8	9	10
	0.25	11	12	13	14	15
	0.5	16	17	18	19	20

Table 2. Population Ability Descriptive Statistics

Condition	**Mean**	**SD**	**Skewness**	**Kurtosis**
1	-0.002730	1.038479	-0.150421	0.188662
2	0.013137	1.029892	-0.069195	0.187633
3	0.025546	1.010727	-0.028604	0.001110
4	0.048284	1.005603	-0.009746	0.026154
5	0.048108	1.000159	-0.029994	0.008582
6	0.009268	1.077265	-0.222996	0.351131
7	0.031122	1.062435	-0.147435	0.103900
8	0.047339	1.028688	-0.077245	0.074921
9	0.080420	1.013663	-0.098288	0.064407
10	0.105576	1.002670	-0.008999	-0.003690
11	0.005810	1.245227	-0.373149	0.114772
12	0.050811	1.185630	-0.310278	0.150854
13	0.128736	1.127389	-0.228015	0.147986
14	0.176984	1.089270	-0.171816	0.121146
15	0.257999	1.048416	-0.157608	0.095907
16	0.001765	1.645387	-0.187875	-0.747664
17	0.114590	1.537622	-0.206840	-0.663610
18	0.247224	1.457308	-0.210405	-0.529366
19	0.382863	1.355198	-0.217774	-0.417714
20	0.502713	1.280387	-0.257944	-0.280208

Table 3. Mean-Shift Ability Distribution Descriptive Statistics

Condition	**Mean**	**SD**	**Skewness**	**Kurtosis**
1	-0.002730	1.038479	-0.150421	0.188662
2	0.013137	1.029892	-0.069195	0.187633
3	0.025546	1.010727	-0.028604	0.001110
4	0.048284	1.005603	-0.009746	0.026154
5	0.048108	1.000159	-0.029994	0.008582
6	0.009268	1.077265	-0.222996	0.351131
7	0.031122	1.062435	-0.147435	0.103900
8	0.047339	1.028688	-0.077245	0.074921
9	0.080420	1.013663	-0.098288	0.064407
10	0.105576	1.002670	-0.008999	-0.003690
11	0.005810	1.245227	-0.373149	0.114772
12	0.050811	1.185630	-0.310278	0.150854
13	0.128736	1.127389	-0.228015	0.147986
14	0.176984	1.089270	-0.171816	0.121146
15	0.257999	1.048416	-0.157608	0.095907
16	0.001765	1.645387	-0.187875	-0.747664
17	0.114590	1.537622	-0.206840	-0.663610
18	0.247224	1.457308	-0.210405	-0.529366
19	0.382863	1.355198	-0.217774	-0.417714
20	0.502713	1.280387	-0.257944	-0.280208

Table 4. Ability Estimate Distribution Descriptive Statistics

Condition	Mean	SD	Skewness	Kurtosis
1	0.063121447	0.754177309	0.581944054	0.192553008
2	0.060135176	0.756830592	0.637256689	0.376974772
3	0.05881665	0.756108023	0.623943927	0.275521334
4	0.054296326	0.756656194	0.625826578	0.296250463
5	0.051826328	0.756487277	0.607845892	0.255848504
6	0.058934386	0.757845068	0.576006362	0.253694526
7	0.052553151	0.758377864	0.574222016	0.192279165
8	0.053472224	0.757651502	0.603069568	0.246487468
9	0.042950453	0.757294558	0.563642212	0.12220677
10	0.036717921	0.760312044	0.605600673	0.2143973
11	0.046035225	0.766810348	0.455938082	-0.157590999
12	0.040023999	0.765660234	0.484664643	-0.057543351
13	0.026333476	0.768197704	0.51170966	0.078754968
14	0.014851657	0.768250312	0.518789633	0.077880201
15	-0.006329743	0.769096494	0.485023578	0.043345578
16	0.016138247	0.818901339	0.46307724	-0.665810477
17	0.008838046	0.812695132	0.425446384	-0.645489508
18	-0.004423298	0.811260566	0.421569041	-0.549104919
19	-0.028872479	0.806533237	0.392846711	-0.444049082

20	-0.049599957	0.801795801	0.3430849	-0.4049522

Table 5. Mean-Shift Ability Estimates Distribution Descriptive Statistics

Condition	Mean	SD	Skewness	Kurtosis
1	0.062310375	0.754449057	0.576302804	0.166374299
2	-0.005379136	0.7687656	0.50014477	0.058447948
3	-0.060428414	0.783986752	0.473278573	0.065088189
4	-0.119573144	0.799889727	0.409702325	0.059513654
5	-0.1691386	0.812113883	0.367179791	0.049879919
6	0.059790297	0.756802159	0.556422546	0.137574381
7	-0.004479892	0.772261772	0.487215042	0.075953912
8	-0.064495457	0.787262953	0.401040756	-0.005253089
9	-0.113881554	0.802355304	0.379742443	0.020196486
10	-0.163588536	0.654675638	0.140087008	0.045748949
11	0.048753725	0.765321053	0.425043796	-0.181195555
12	-0.004522168	0.781715647	0.38460532	-0.222546701
13	-0.053850017	0.795782455	0.315809886	-0.218999559
14	-0.101921035	0.814393909	0.261439671	-0.244087594
15	-0.149478962	0.838240215	0.263274456	-0.121425386
16	0.017787584	0.818126748	0.460878949	-0.697815518
17	-0.021167916	0.838759543	0.419721942	-0.698605252
18	-0.044206006	0.862599796	0.420603576	-0.617895676
19	-0.058785996	0.885664848	0.415195485	-0.550136309
20	-0.06502776	0.921186008	0.429914387	-0.486235603

Table 6. Ability Estimates Mean difference (subgroup, majority group)

Subgroup Proportion of Population		Subgroup Ability Expected Growth			
0.05 (1000:19000)	Sub group expected mean growth	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean growth	0.1212	0.2401	0.3914	0.5461
	Majority group estimated mean changes	-0.0095	-0.0172	-0.0299	-0.0406
0.1 (2000:18000)	Sub group expected mean growth	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean growth	0.0896	0.2164	0.3181	0.4829
	Majority group estimated mean changes	-0.0170	-0.0301	-0.0531	-0.0783
0.25 (5000: 15000)	Sub group expected mean growth	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean growth	0.0626	0.1221	0.1976	0.2879
	Majority group estimated mean changes	-0.0289	-0.0670	-0.1074	-0.1658
0.5 (10000:10000)	Sub group expected mean growth	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean growth	0.0221	0.0395	0.0578	0.0803
	Majority group estimated mean changes	-0.0367	-0.0806	-0.1478	-0.2118

Table 7. Mean-Shift Ability Estimates Mean difference (subgroup, majority group)

Subgroup Proportion of Population	Subgroup theta mean after equating Growth				
0.05 (1000:19000)	Sub group mean difference	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean difference	0.1191	0.1433	0.2731	0.3403
	Majority group mean changes	0.0650	0.1217	0.1771	0.2257
0.1 (2000:18000)	Sub group mean difference	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean difference	0.0656	0.1754	0.2341	0.3165
	Majority group mean changes	0.0641	0.1186	0.1670	0.2130
0.25 (5000: 15000)	Sub group mean difference	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean difference	0.0626	0.1399	0.2171	0.2876
	Majority group mean changes	0.0502	0.0902	0.1285	0.1684
0.5 (10000:10000)	Sub group mean difference	**0.25**	**0.50**	**0.75**	**1.00**
	Subgroup estimated mean difference	0.0563	0.0965	0.1304	0.1554
	Majority group mean changes	0.0216	0.0275	0.0227	0.0103

Table 8. Condition 1 Total Population Classification Contingency Table

	Estimation Class 1	Estimation Class 2	Estimation Class 3	Estimation Class 4	Classification based on ability
Ability Class 1	5471	1049	0	0	**6520**
Ability Class 2	0	3290	19	0	**3309**
Ability Class 3	0	256	7164	0	**7420**
Ability Class 4	0	0	862	1889	**2751**
Classification based on Estimation	**5471**	**4595**	**8045**	**1889**	**20000**

Table 9. Total Population Over-estimation/Under-estimation Proportion

Subgroup Proportion of Population		Subgroup Ability Expected Growth				
0.05 (1000:19000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0534	0.0459	0.03975	0.03025	0.02825
	Under-estimation Proportion	0.0559	0.06475	0.07075	0.08665	0.08275
0.1 (2000:18000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.04585	0.03925	0.0336	0.0141	0.0058
	Under-estimation Proportion	0.07655	0.0961	0.09375	0.11495	0.1377
0.25 (5000: 15000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.04005	0.02745	0.0019	0.00005	0.0000
	Under-estimation Proportion	0.1263	0.1446	0.19025	0.2386	0.3439
0.5 (10000:10000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.02825	0.006	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.1836	0.22905	0.3131	0.44025	0.5276

Table 10. Subgroup Over-estimation/Under-estimation Proportion

Subgroup Proportion of Population		Subgroup Ability Expected Growth				
0.05 (1000:19000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0450	0.0440	0.0430	0.0290	0.0290
	Under-estimation Proportion	0.0120	0.0220	0.0330	0.0640	0.0680
0.1 (2000:18000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0390	0.0430	0.0355	0.0120	0.0080
	Under-estimation Proportion	0.0200	0.0305	0.0475	0.0660	0.1070
0.25 (5000: 15000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0460	0.0312	0.0032	0.0002	0.0000
	Under-estimation Proportion	0.0288	0.0492	0.1012	0.1504	0.2830
0.5 (10000:10000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0480	0.0102	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.0593	0.1109	0.2150	0.3736	0.4780

Table 11. Majority Group Over-estimation/Under-estimation Proportion

Subgroup Proportion of Population		Subgroup Ability Expected Growth				
0.05 (1000:19000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0538	0.0460	0.0396	0.0303	0.0282
	Under-estimation Proportion	0.0582	0.0670	0.0727	0.0878	0.0835
0.1 (2000:18000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0466	0.0388	0.0334	0.0143	0.0056
	Under-estimation Proportion	0.0828	0.1034	0.0989	0.1204	0.1411
0.25 (5000: 15000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0381	0.0262	0.0015	0.0000	0.0000
	Under-estimation Proportion	0.1588	0.1764	0.2199	0.2680	0.3642
0.5 (10000:10000)	Sub group expected mean growth	**0.00**	**0.25**	**0.50**	**0.75**	**1.00**
	Over-estimation Proportion	0.0085	0.0018	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.3079	0.3472	0.4112	0.5069	0.5772

Table 12. Mean-Shift Total Population Over-estimation/Under-estimation Proportion

Subgroup Proportion of Population		Subgroup Ability Expected Growth				
0.05 (1000:19000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0488	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.0674	0.3304	0.5798	0.7565	0.8522
0.1 (2000:18000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0562	0.0000	0.0000	0.0000	0.0483
	Under-estimation Proportion	0.0734	0.3360	0.5816	0.7463	0.7287
0.25 (5000: 15000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0451	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.1251	0.3458	0.5704	0.7226	0.7661
0.5 (10000:10000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0278	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.1771	0.3153	0.4397	0.5003	0.5492

Table 13. Mean-Shift Subgroup Over-estimation/Under-estimation Proportion

Subgroup Proportion of Population		Subgroup Ability Expected Growth				
0.05 (1000:19000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0360	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.0110	0.1520	0.3180	0.4390	0.5550
0.1 (2000:18000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0555	0.0000	0.0000	0.0000	0.0430
	Under-estimation Proportion	0.0190	0.1515	0.3235	0.4550	0.5195
0.25 (5000: 15000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0496	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.0264	0.1520	0.3222	0.4564	0.5426
0.5 (10000:10000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0462	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.0595	0.1940	0.3557	0.4579	0.5502

Table 14. Mean-Shift Majority Group Over-estimation/Under-estimation Proportion

Subgroup Proportion of Population		Subgroup Ability Expected Growth				
0.05 (1000:19000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0494	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.0704	0.3398	0.5936	0.7732	0.8678
0.1 (2000:18000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0562	0.0000	0.0000	0.0000	0.0488
	Under-estimation Proportion	0.0794	0.3564	0.6103	0.7787	0.7519
0.25 (5000: 15000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0435	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.1579	0.4103	0.6531	0.8113	0.8406
0.5 (10000:10000)	Sub group expected mean growth	**0**	**0.25**	**0.5**	**0.75**	**1**
	Over-estimation Proportion	0.0093	0.0000	0.0000	0.0000	0.0000
	Under-estimation Proportion	0.2947	0.4365	0.5237	0.5427	0.5481

APPENDIX B: FIGURES

Figure 1. Subgroup/Major group Population Ability Density Distribution-Condition 1-5

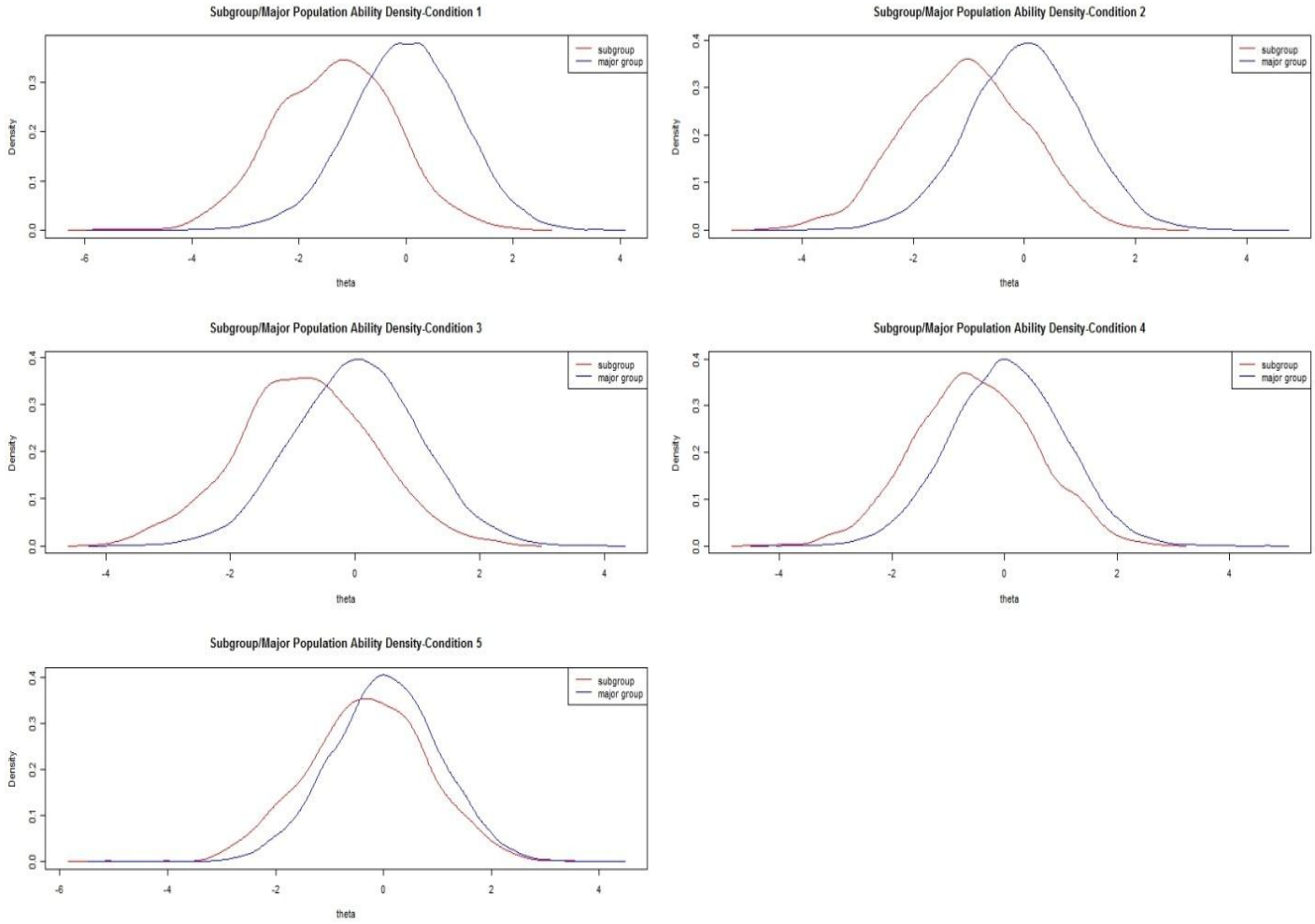

Figure 2. Population Ability Density Distribution-Condition 1-20

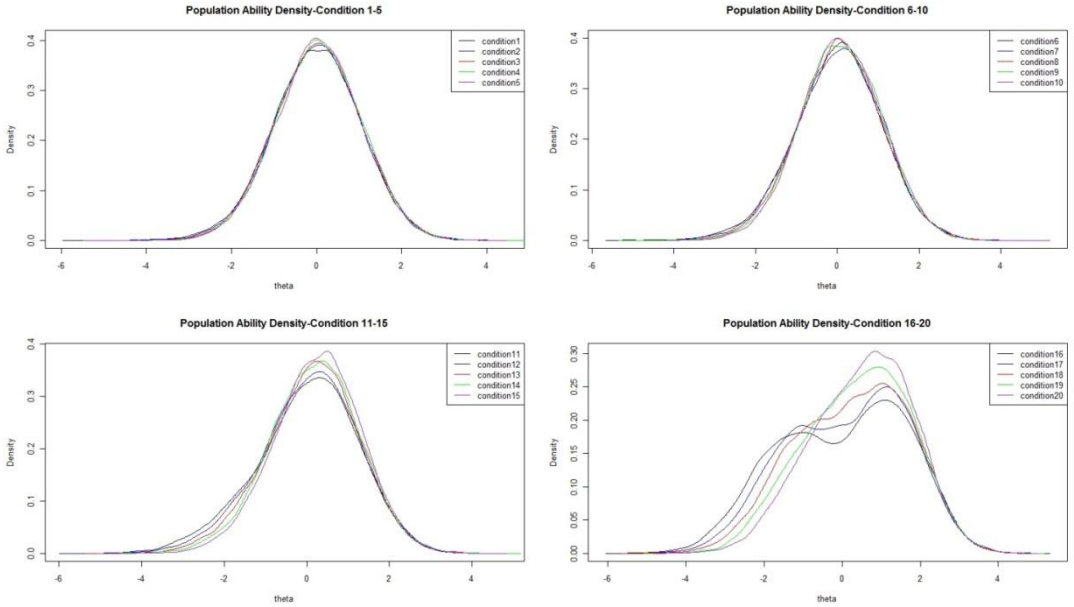

Figure 3. Mean-Shift Ability Density Distribution-Condition 1-20

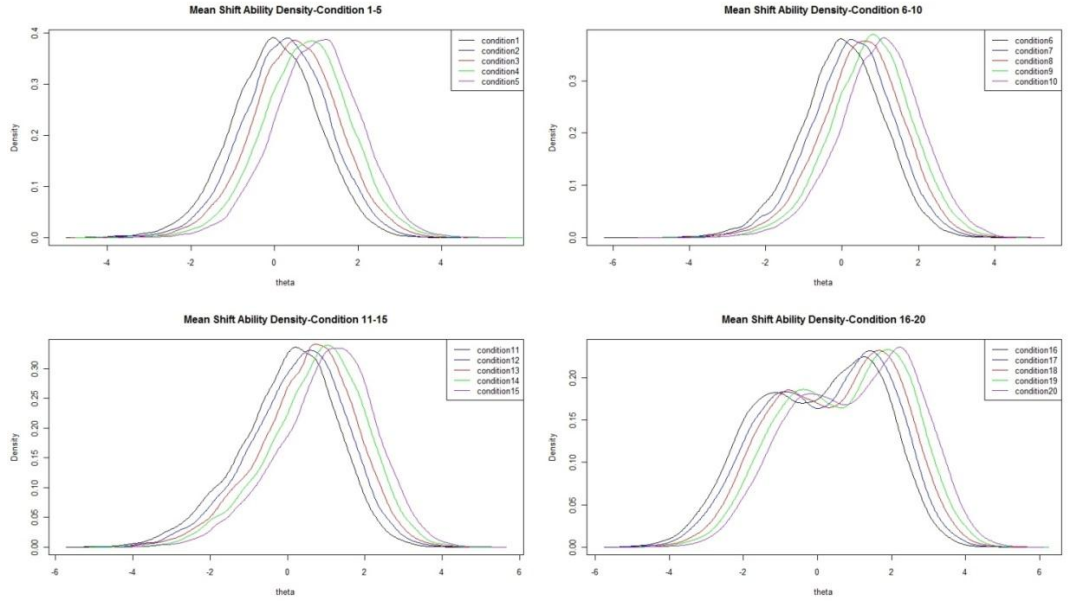

Figure 4. Ability Estimates Density Distribution Condition 1-20

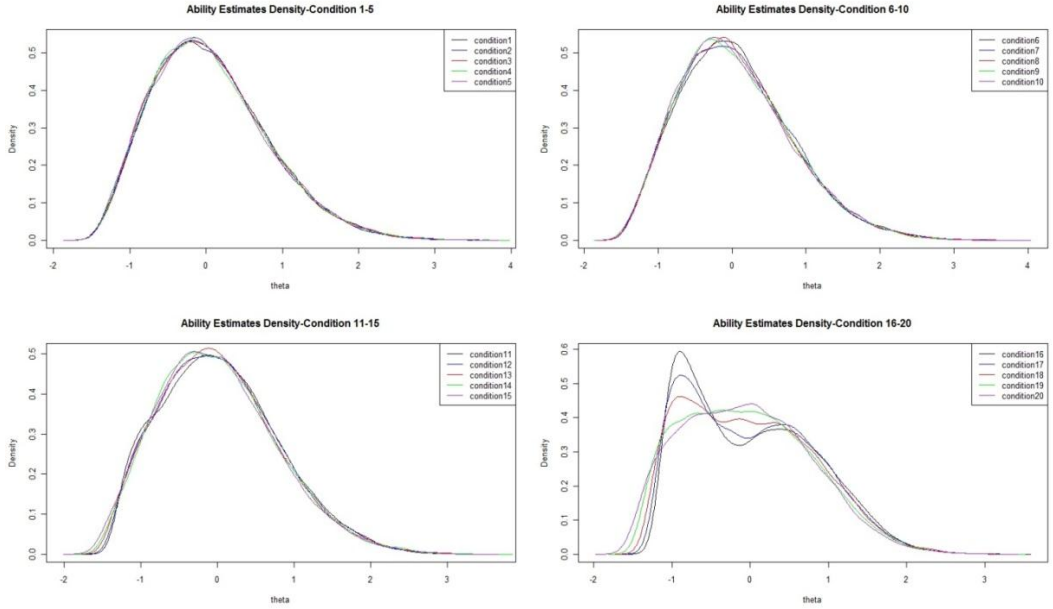

Figure 5. Mean-Shift Ability Estimates Density Distribution Condition 1-20

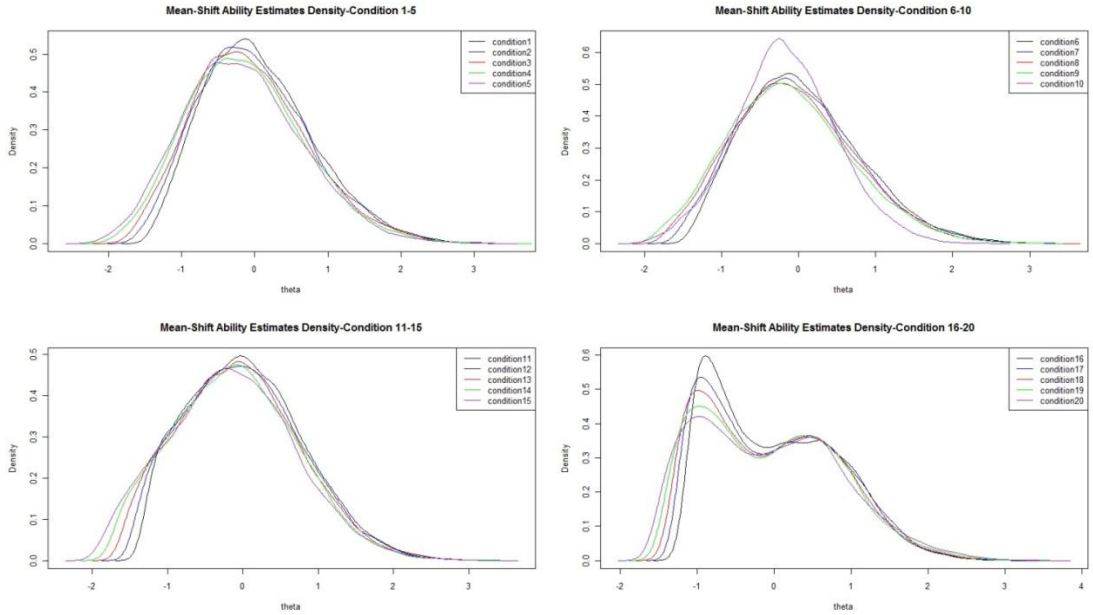

Figure 6. Test Characteristic Curve Example