Psychological Test and Assessment Modeling, Volume 53, 2011 (4), 499-522

Exploring levels of performance using the mixture Rasch model for standard setting[1]

Hong Jiao[2], Robert W. Lissitz[3], George Macready[3], Shudong Wang[4] & Senfeng Liang[3]

Abstract

This study proposes a method using the mixture Rasch model to set performance standards. The mixture Rasch model classifies examinees into qualitatively distinct latent groups based on the information in the response patterns and quantifies individual differences within-group utilizing a continuous latent trait.The goals of this study are twofold. One is to demonstrate an application of the mixture Rasch model to statistically validate the performance proficiency levels set by policy makers and reviewed by content experts. The other is to demonstrate how performance cut scores can be obtained based on the results from data analysis using the mixture Rasch model. In general, this study presents a mixture Rasch model-based approach for setting performance standards which is expected to facilitate the standard setting process by providing the data driven information related to the policy specified performance levels and the performance cut scores set by the panelists.

Key words: mixture Rasch model, standard setting, proficiency classification, model-based classification, performance standards

[1] An earlier version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education in Denver, Colorado (Jiao, Lissitz, Macready, Wang, & Liang, 2010).

[2] *Correspondence concerning this article should be addressed to:* Hong Jiao, PhD, Department of Measurement, Statistics and Evaluation, 1230B Benjamin Building, University of Maryland, College Park, MD 20742, USA; email: hjiao@umd.edu

[3] University of Maryland

[4] NWEA

Classifying examinees into different proficiency levels is a frequently expected outcome from various testing programs. In licensure and certification testing programs, classification of examinees is required to select qualified candidates in various professional fields. In educational testing, classification of students into different proficiency categories serves the purpose of screening students for instruction, grade promotion, selection, or admission. In counseling, classification is made to distinguish people with different trait or attitude levels which allows for different treatments being advised. In language assessment, classifying examinees into different levels of language proficiency is needed for a wide range of purposes such as selection, task assignments, or instructional decisions. In general, classification is needed to categorize examinees into distinct groups in terms of the trait the instrument is intended to measure.

A classification decision is made by determining whether an examinee has reached the minimal level of competency for a certain category among the dichotomous or polytomous categories. A dichotomous decision often refers to a decision made by classifying an examinee into one of two mutually exclusive and exhaustive categories such as pass/fail, or mastery/non-mastery, or qualified/unqualified, or proficient/non-proficient. While polytomous decisions simply indicate that the number of categories within which examinees are classified is three or greater such as fail/pass/proficient/excellent. These different ordered proficiency levels are delimited and defined by the cut scores on the assessed latent trait of interest.

Standard setting is a procedure used to set one or more cut scores indicating the minimal level of competence required to classify an examinee into one of the performance levels. Standard setting methods can be classified into two major categories (Jaeger, 1993; Lau, 1996; Stephenson, et al. 2000). One category is test-centered while the other is examinee-centered (Jaeger, 1993).

The test-centered methods are based on judgments about test items and include the Nedelsky method (Nedelsky, 1954), the Angoff method (Angoff, 1971) and variants of the Angoff method (the modified Angoff methods; Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997), the Ebel method (Ebel, 1972), the Jaeger method (Jaeger, 1982), the bookmark method (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001), and the body of work method (Kahl, Crockett, DePascale, & Rindfleisch, 1994, 1995; Kingston, Kahl, Sweeney & Bay, 2001). Currently the most widely used methods are the modified Angoff method, the bookmark method, and the body of work method for performance items. In the application of the Angoff method to set a cut score, judges are asked to estimate the probability of an examinee with the minimal competence answering an item correctly or the number of borderline students out of 100 who would be expected to get the item correct. The sum of the probabilities for all items or the proportion of borderline students answering items correctly will be a judge's estimate of the total score of a minimally competent examinee. The sum of each judge's scores will be the cut score under this approach. The modified Angoff methods were proposed to improve the Angoff method. When using the bookmark method for standard setting, items are rank ordered in terms of difficulty from the easiest to the most difficult. Item difficulty estimates are obtained from item response theory (IRT) calibration. Panelists are asked to find the location in the ordered item list that separates examinees into categories. The body of work method is intended for tests using open-ended items like writing assessments. Panelists are required to read and sort students' work into different performance categories. Obviously, judges and the judges' conceptualization of minimally competent examinees, and judges' knowledge about the content of each test item will greatly affect the standards set using the test-centered methods.

The examinee-centered methods are based on judgments about individual examinees. There are three commonly used methods within this class, the contrasting groups method (Zieky & Livingston, 1977), the borderline group method (Zieky & Livingston, 1977), and the up and down method (Livingston & Zieky, 1982). The contrasting groups method divides examinees into three groups: competent, borderline, and incompetent. The standard is set based on the analyses of the test score distributions on examinees in the competent and incompetent groups excluding those examinees in the borderline group. Hambleton & Eignor (1980) suggested using the intersection point of the test score distributions for the competent and incompetent groups as the standard. Livingston & Zieky (1982) suggested computing the percentage of examinees classified as competent for each test score. The standard is the score where 50% of the examinees are classified as competent. The borderline-group method requires that judges be able to determine the requisite level of knowledge or skill a competent examinee possesses. Examinees are grouped into three categories: competent, borderline, and incompetent. Those identified as borderline examinees take the competency test. The cut score is the median of the distribution of test scores of borderline examinees. The up and down method is a variation of the contrasting groups method. An examinee with a test score near where the passing score is predicted to be is identified and this examinee's competency is judged. If the first examinee is judged qualified, a second examinee with a test score lower than this examinee is selected and the second examinee's competency is judged. If the first examinee is judged unqualified, a second examinee with a test score higher than this examinee is selected and the second examinee's competency is judged. This process continues by choosing examinees based on the judgment of the previously selected examinees. This method moves down from test score levels where examinees are qualified and up from test scores where examinees are unqualified. The cut score is set at a level where an examinee is about as likely to be qualified as to be unqualified. The successful implementation of the examinee-centered methods relies on the correct identification of examinees in the competent, incompetent, and/or borderline groups.

The selection of a standard setting method is vital in setting performance cut scores. The stability of the cut score set by a particular method depends on the features of the population of judges and the simplicity of implementing a method. Jaeger (1993) and Stephenson et al. (2000) summarized the comparability of the results obtained using different standard setting methods. The differences in the cut score values resulting from using different standard setting methods are not negligible. Hambleton (1980), Koffler (1980), and Shepard (1980, 1984) suggested that it might be better to use several methods in any study and consider all results with other non-statistical factors to determine a cut score. The main reason for this instability (and inconsistency across methods) of cut scores set within a standard setting method and across standard setting methods is the subjectivity inherent in these procedures.

All these current widely used approaches for standard setting are based on the judgmental processes of collecting empirical evidence regarding the unobserved performance level of examinees based on panelists' knowledge of the test content, target examinee population and item information. However, these judgmental processes are subject to criticisms due to the subjectivity in evaluating or estimating examinees' performance on item samples. Judgmental errors may result from such factors as the substantiation of a minimally competent examinee or panelists' ability to estimate the probability that a minimally competent examinee would answer an item correctly. In addition, the number of performance categories is a policy decision (Hambleton & Pitoniak, 2006). Whether or not the policy defined performance levels bear practical meaning, there is no step in the standard setting process that allows for verification. The predetermined performance levels may only be a proxy of the true latent proficiency classes of the examinee population and may not effectively match the number of the latent performance levels. Moreover, current standard setting methods only make use of the information in the sum score or item difficulty, while additional information provided by the response patterns is not utilized in the classification process. This is epitomized by the ordered item booklet in the bookmark method (Lewis et al., 1996; Mitzel et al., 2001) guiding panelists through the items ranked by difficulty. The body of work (Kahl et al., 1994, 1995; Kingston et al., 2001) method uses pinpointing and range finding both of which focus on the scores of the students on the test materials. The Modified Angoff method (Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997) also focuses on the estimated performance by a minimally competent test taker on each individual item. Again, none of these methods focus on respondents' patterns of item performance.

To reduce the level of subjectivity built into the current common practice of standard setting and to make full use of the information in item response data, some researchers have explored using model based approaches for standard setting. These approaches include latent class modeling (Brown, 2000; Luecht & DeChamplain, 1998; Templin, Poggio, Irwin, & Henson, 2007), attribute hierarchy modeling (Sadesky & Gushta, 2004), and cluster analysis (Sireci, 1995, 2001; Sireci, Robin, & Patelis, 1999). Under these model based approaches for standard setting, less human judgment is involved. Instead, the classifications of the examinees are based on the model that is fitted to the test data. However, the models used in these model based approaches can identify the latent groups but not estimate the latent ability simultaneously. Due to this limitation, these methods use a two-stage approach to classify examinees into groups and then an approximation to the cut scores is set based upon raw scores or latent ability estimates from an IRT model.

This paper intends to explore potentially more efficient and justifiable standard setting procedures (e.g., Lissitz and Kroopnick, 2007) based on latent variable models by reducing the limitations of the judgmental standard setting approaches and obtaining cut scores on the latent ability scale that are based on a model based analysis. The proposed procedure is demonstrated using a latent variable model, the mixture Rasch model, to identify latent classes based on the item response data and then analytically solving equations to find the intersecting point of two adjacent distributions which are used as the cut point for distinguishing between the two adjacent latent classes. In this paper we explore using the mixture Rasch model which models a mixture of a discrete latent class dimension and a continuously measured trait in describing the qualitative characteristics of the latent classes and the quantitative within-group and between-group characteristics. We expect that this approach facilitates identification of borderline/minimally competent examinees.

The mixture Rasch model

The mixture Rasch model was first proposed by Kelderman and Macready (1990), Mislevy and Verhelst (1990) and Rost (1990) to model the test data with more than one latent population by integrating the Rasch measurement model (Rasch, 1960), where a continuous latent variable underlies the performance of examinees, and a latent class model, where the class membership underlies the performance of examinees, in responding to items. The mixture Rasch model assumes that examinees are from multiple latent populations and the Rasch model holds within each latent class/population with unique item difficulty parameters differing across various latent classes. The combination of the Rasch model and the latent class model allows for simultaneous estimation of a continuous latent ability as well as latent group membership. The latent class membership for each examinee is not observed but estimated based on the information provided by the item response patterns. Thus each examinee is characterized by two latent variables, a continuous quantitative variable (that provides a measure of the quantitative trait of interest) and a categorical qualitative variable (which diffentiates among respondents who differ in their likelihood of correctly responding to items). The quantitative variable is analogous to the latent ability parameter estimates in the Rasch model with the same interpretation. The qualitative variable is a categorical variable indicating the latent class membership. This discrete variable is an underlying variable also influencing the performance of an examinee to a certain item. Thus, an examinee's performance on an item is determined by its discrete qualitative group membership and the continuous quantitative latent ability.

In the mixture Rasch model, the probability of a correct response for person j to item i conditional on the person's latent class membership g is expressed in equation 1.

$$p_{jig} = \frac{1}{1+\exp[-(\theta_{jg} - b_{ig})]}, \tag{1}$$

where P_{jig} is the probability of the j^{th} examinee with a latent ability of θ_{jg} in latent class g responding correctly to the i^{th} item with difficulty b_{ig} for that particular latent class g. For every examinee, θ_{jg} is a continuous ability estimate. In addition, each examinee's group identity is estimated to be g and labeled in the estimated theta as $\theta_{jg.}$ Each examinee's latent group membership is determined by comparing the posterior probability of that particular examinee being in each latent group. The assignment of examinees to a latent group is based on the magnitude of the posterior probabilities for their respective response patterns. Examinees are assigned to that group for which their posterior probability is the largest.

The unconditional probability of a correct response is expressed in equation 2.

$$p_{ji} = \sum_g \pi_g p_{jig} = \sum_g \pi_g \frac{1}{1+\exp[-(\theta_{jg} - b_{ig})]}, \tag{2}$$

where P_{ji} is the unconditional response probability and π_g is the class mixing proportion; with constraints 0 < π_g < 1 and $\sum_g \pi_g = 1$ across classes.

The mixture Rasch model has been applied in solving various psychometric problems. For instance, differential item parameter estimates across latent classes have been utilized in identifying differential item functioning (DIF) across the latent groups of examinees (Cohen & Bolt, 2005; Dai & Mislevy, 2006; De Ayala et al., 2002; Kelderman & Macready, 1990; Lu & Jiao, 2009; Samuelson, 2005). Different item response patterns were utilized to identify latent classes utilizing different cognitive strategies solving problems (Mislevy & Verhelst, 1990; Rijmen & De Boeck, 2003; Rost & von Davier, 1993). Different latent subpopulations caused by test speededness can be identified by data analysis using the mixture Rasch model (Bolt, Cohen, & Wollack, 2002; Boughton & Yamamoto, 2007; Yamamoto, 1989; Yamamoto & Everson, 1997).

Validating the number of performance levels and finding cut scores

This paper creates a framework for standard setting using results from the mixture Rasch model. The current practice of standard setting relies on human judgments by either policy makers or other test stakeholders to identify the number of proficiency levels which may be essentially different latent groups or populations. Then content experts are asked to conceptualize and/or delineate the features of examinees in different groups such as minimally competent, competent, incompetent or borderline depending on the method used for standard setting (test-centered or examinee-centered). Once the cut scores are established, they are applied to classify examinees into various proficiency categories.

In standard setting, policy makers make decisions regarding the number of proficiency levels in the student population. If the number of proficiency levels is too large, the examinees in some proficiency levels may be essentially the same as those in adjacent groups in terms of their academic achievement. If the number of proficiency levels is too small, the examinees' characteristics in one proficiency level may be too heterogeneous. Thus homogeneous descriptions of the examinees' characteristics in that specific proficiency level may not be sufficiently accurate. The purpose of standard setting is to use human judges to identify distinct groups in terms of the latent ability which is estimated using the information in the item responses. After classifying the examinees into different proficiency groups, it is valid to assume that the examinees in the same proficiency level perform more similarly to each other than to examinees in other proficiency categories. Thus, the response patterns observed for examinees in the same proficiency level resemble each other. This is exactly what the estimates from the mixture Rasch model are based on. In applying the mixture Rasch model, the latent groups will be identified based on the information in the item response patterns. The identification of the latent groups based on model fit indices may be used as a validation measure of the performance categories set by policy makers. The group indicator in equation 1 for every examinee will be assigned based on the largest posterior probability of being in each group. Simultaneously, a continuous theta is estimated to represent the relative standing of examinees within groups. The use of the mixture Rasch model fulfills the tasks in the conventional two-stage standard setting process in an one-step model based analysis. The implementation is as follows.

To start the mixture Rasch model based standard setting process, a test form constructed conforming to the test specification is administered to a large and representative sample of the examinee population. The use of a large number of examinees is needed to accurately capture the possible latent groups that may be present in the population of interest. After the test administration, test data are analyzed using the mixture Rasch model. Multiple estimation methods are available (Liu & Jiao, 2010). These include the marginal maximum likelihood estimation (MMLE) method with the expectation-maximization (EM) algorithm implemented in the Multidimensional Discrete Latent Trait Model (mdltm) software (von Davier, 2005) and M-Plus (Muthen & Muthen, 2007), the conditional maximum likelihood estimation method in the Winmira software (von Davier, 1994), and the Markov Chain Monte Carlo estimation method in the WINBUGS program (Bolt, Cohen, & Wollack, 2002; Cohen & Bolt, 2005; Samuelson, 2005). This paper demonstrates the procedure using the estimates from the MMLE method with the EM algorithm implemented in the mdltm software (von Davier, 2005).

The mdltm software (von Davier, 2005) outputs the estimated mean and standard deviation for each estimated latent class. In this application, we assume the latent trait distribution in each group to be normally distributed. Following the suggestions by Hambleton and Eignor (1980) using the intersection point of the test score distributions for the more competent and less competent groups as the standard. The intersecting point of the two adjacent latent class distributions, as illustrated in the plot in Figure 1, may be used as the cut score on the continuous latent ability scale to distinguish the two adjacent classes with a certain degree of error which can be computed based on the intersecting points on the density fund for the two latent distributions. Note that under this approach for specifying cut scores the estimated likelihood of classification error is minimized.

The analytic solution for finding the intersection point of the two adjacent classes is demonstrated as follows. The estimated latent classes which are assumed to be normally distributed are treated as the latent proficiency levels. The function of a normal distribution is represented by equation 3 as follows:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad . \tag{3}$$

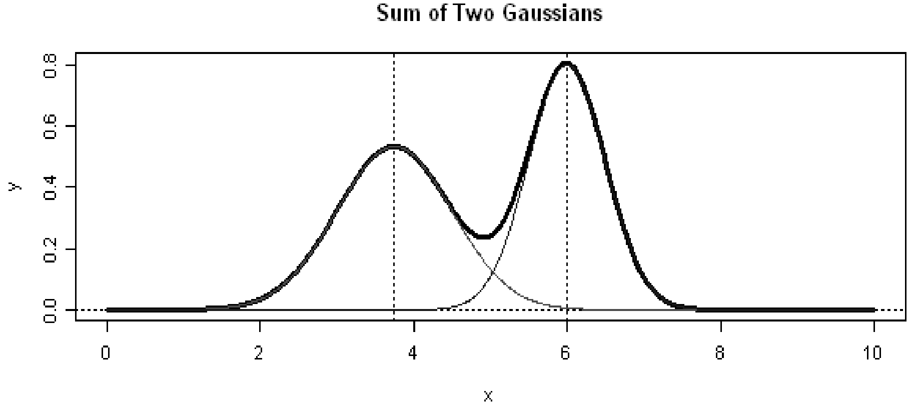


Figure 1:
The mixture of two normal distributions

Assume that there is a mixture of *J* normal distributions each defined as: $N(u_j,\sigma_j)$, then the intersecting points of two adjacent normal distributions *j* and *j +1* (as shown in Figure 1) can be found by equating the density functions for these two adjacent distributions and solving for *X* seen in equation 4:

$$w_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} = w_{j+1} \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{j+1}}} e^{-\frac{(x-\mu_{\mathrm{j+1}})^2}{2\sigma_{\mathrm{j+1}}^2}} , \tag{4}$$

where w_{j} is the latent class proportion corresponding to the *j*th latent class and w_{j+1} is the latent class proportion corresponding to the *j*+1 latent class.

Assume that 5.9% of the total population is in the jth latent class, $N(u_j,\sigma_j)$, and 13.2% is in the *j* + 1 latent class, $N(u_{j+1},\sigma_{j+1})$, with $\mu_j < \mu_{j+1}$. Then we weight $N(u_j,\sigma_j)$ by $w_j = 0.059$ and $N(u_{j+1},\sigma_{j+1})$ by $w_{j+1} = 0.132$. Details related to finding the value of *x* that corresponds to the intersecting point for adjacent distributions on the latent ability scale with unequal mixing proportions can be found in Appendix A.

When multiple latent classes are estimated, the intersecting points for each pair of adjacent distributions can be computed based on equation 4 by inserting the estimated mean, standard deviation, and the mixing proportion for each distribution. Since this is a quadratic equation, there are two roots. Based on the relative position of each distribution, the appropriate root may be identified. The resulting value of *X* corresponds to the intersecting point for the two distributions within the commonly used logit theta scale ranging from -4 to +4.

Methods

To illustrate the procedure of using the mixture Rasch model for standard setting, this section uses simulation data based on the standard setting results from a large-scale assessment using the bookmark method. The simulated large-scale assessment is a language proficiency test. It consists of assessments of multiple subskills including reading and listening. Only reading data are used for the purpose of illustration. There were five proficiency levels specified for the reading test. The cut scores on the theta scale were -1.8, -0.6, 0.96, and 1.68 with 6%, 13.2%, 52.6%, 22.3%, and 5.9% of the students falling into each of the five proficiency levels.

The simulation data were generated under the assumption that examinees in different proficiency levels/classes were from distinct normal distributions. It was further assumed that the midpoint of the intersecting points was the mean of the distribution and the intersecting points were two standard deviations away from the mean. Based on the theta cuts obtained from the large-scale reading assessments, the targeted true distributions for the five simulated proficiency levels were computed as shown in Table 1. The mean ability increased from -2.46 to 2.58 from proficiency level 1 to 5. The middle level of proficiency had the largest proportion and the two extreme proficiency levels: level 1 and level 5 had the smallest proportion. The generated true ability distribution for each latent group was close to the targeted true values as shown in Table 1.

To make sure that the sample size for each group was not a potential problem for model parameter estimation, this study simulated 10,000 examinees so that the sample size was adequate for extreme high or low proficiency levels with a small mixing proportion around 6%. Thus, the sample size for each group from the lowest to the highest proficiency level was 600, 1300, 5300, 2200, and 600.

The item parameters were generated to reflect the differences among the five proficiency groups. The item parameters for the middle proficiency level, level 3, were simulated from a standard normal distribution with a mean of 0 and a standard deviation of 1. The item parameters for other proficiency levels were generated by adding or subtracting

Table 1:
The target and the generated ability distribution for different proficiency levels

	Targeted True Values			**Generated True Values**		
Proficiency Levels	Mean	Standard Deviation	Mixing Proportion	Mean	Standard Deviation	Mixing Proportion
1	-2.46	0.315	6%	-2.4624	0.3296	6%
2	-1.20	0.30	13%	-1.1989	0.2974	13%
3	0.18	0.39	53%	0.1827	0.3934	53%
4	1.32	0.18	22%	1.3224	0.1804	22%
5	2.58	0.315	6%	2.5943	0.3281	6%

some constants to the generated item parameters for level 3. In general, most of the items for less proficient groups (level 1 and level 2) are more difficult while most of the items for more proficient groups (level 4 and level 5) are easier. However, to maintain the constraints of $\sum_i b_{ig} = 0$ within each latent group, a small subset of items need to be manipulated following the counter intuitive trend. That is, a small portion of the items are easier for the less proficient groups and another small portion of items are more difficult for the more proficient groups. This study simulated 40 items, a commonly used test length in large-scale assessments. The generated item parameters are presented in Table 2. The magnitudes and patterns of differences in item difficulties simulated in our study are similar to studies related to the mixture Rasch model. For instance, in Li et al. (2009), the largest difference in item parameters was 3 which is greater than those simulated in our study with the largest difference of 2. In Rost (1990), the largest item difficulty parameter difference was 5.4.

After the generation of true item and person ability parameters for each proficiency level, the item responses were generated using these true model parameters in equation 1. Then a series of mixture Rasch models with differing numbers of latent classes: 1, 2, 3, 4, 5, 6, and 7 were fitted to the response data using the mdltm software. Model convergence was checked and model fit was evaluated using multiple fit indexes from the mdltm software. Once the best fitting mixture Rasch model was identified, the intersecting points were computed and treated as theta cut points for classification decisions. Lastly, the generated item response data were fitted to the Rasch model and the classification decisions were made based on the theta cut points obtained from the proposed procedure based on the mixture Rasch model. These classification decisions were then used to assess the accuracy of classifications in identifying the correct group membership of simulated respondents.

Results

The mixture Rasch models with different numbers of latent classes were fitted to the generated item response data. These include mixture Rasch models with 1 through 7 latent classes. However, the mixture Rasch models with 6 and 7 latent classes did not converge and therefore were eliminated from further consideration. The remaining models resulted in the converged estimations. The model fit based on Akaike's (1974) information criterion (AIC), Schwarz's (1978) Bayesian information criterion (BIC), and deviance are summarized in Table 3. Both the deviance and the AIC supported the selection of a 5-class mixture Rasch model as the preferred model based on its fit to the data. However, the BIC resulted in the 3-class mixture Rasch model being preferred.

To better understand the model data fit, the estimated mean, standard deviation, and mixing proportions were compared for the 5-class and the 3-class mixture Rasch models and are summarized in Table 4. The middle classes, class 3 for the 5-class solution and class 2 for the 3-class solution were generally similar in terms of their means and standard deviations. The mixing proportion for the middle class from the 3-class solution was

Table 2:
Generated item parameters for different proficiency levels

Proficiency Levels				
1	2	3	4	5
-1.86531	-1.86531	-1.86531	-1.86531	-2.86531
-1.41574	-1.41574	-1.41574	-1.41574	-2.41574
-1.39954	-1.39954	-1.39954	-1.39954	-2.39954
-1.17929	-1.17929	-1.17929	-1.17929	-2.17929
-1.13338	-1.13338	-1.13338	-1.13338	-2.13338
-1.10772	-1.10772	-1.10772	-1.10772	-2.10772
-0.82712	-0.82712	-0.82712	-0.82712	-1.82712
-0.72216	-0.72216	-0.72216	-0.72216	-1.72216
-0.70265	-0.70265	-0.70265	-0.70265	-1.70265
-0.64371	-0.64371	-0.64371	-0.64371	-1.64371
-2.54446	-1.54446	-0.54446	-1.04446	-0.54446
-2.54367	-1.54367	-0.54367	-1.04367	-0.54367
-2.40944	-1.40944	-0.40944	-0.90944	-0.40944
-2.39603	-1.39603	-0.39603	-0.89603	-0.39603
-2.38524	-1.38524	-0.38524	-0.88524	-0.38524
-2.31254	-1.31254	-0.31254	-0.81254	-0.31254
-2.24428	-1.24428	-0.24428	-0.74428	-0.24428
-2.21175	-1.21175	-0.21175	-0.71175	-0.21175
-2.11729	-1.11729	-0.11729	-0.61729	-0.11729
-2.10656	-1.10656	-0.10656	-0.60656	-0.10656
0.916372	0.416372	-0.08363	0.916372	1.916372
0.936383	0.436383	-0.06362	0.936383	1.936383
0.954654	0.454654	-0.04535	0.954654	1.954654
1.00308	0.50308	0.00308	1.00308	2.00308
1.006064	0.506064	0.006064	1.006064	2.006064
1.023848	0.523848	0.023848	1.023848	2.023848
1.262404	0.762404	0.262404	1.262404	2.262404
1.352435	0.852435	0.352435	1.352435	2.352435
1.482835	0.982835	0.482835	1.482835	2.482835
1.612247	1.112247	0.612247	1.612247	2.612247
1.617248	1.117248	0.617248	0.117248	-0.38275
1.628146	1.128146	0.628146	0.128146	-0.37185
2.039357	1.539357	1.039357	0.539357	0.039357
2.084255	1.584255	1.084255	0.584255	0.084255
2.414787	1.914787	1.414787	0.914787	0.414787
2.448211	1.948211	1.448211	0.948211	0.448211
2.5045	2.0045	1.5045	1.0045	0.5045
2.63387	2.13387	1.63387	1.13387	0.63387
2.696744	2.196744	1.696744	1.196744	0.696744
2.897149	2.397149	1.897149	1.397149	0.897149

Table 3:
Summary of model fit

	1 class	2 class	3 class	4 class	5 class
Deviance	452864.1	445557.7	443691.7	443286.1	443069.2
AIC	452956.1	445739.7	443963.7	443648.1	443521.2
BIC	453287.8	446395.8	444944.3	444953.2	445150.7

Table 4:
Estimated distributions for the 5-class and 3-class mixture Rasch models

# of Latent Classes	Class	Mean	Standard Deviation	Mixing Proportion
5-class	1	-2.0368	0.3123	6.26%
	2	-0.8563	0.3836	12.57%
	3	0.1228	0.3023	52.86%
	4	0.9553	0.1430	22.11%
	5	2.4069	0.2419	6.19%
3-class	1	-1.297	0.362	15.87%
	2	0.132	0.354	59.22%
	3	1.181	0.318	24.91%

larger than that for the 5-class solution. This indicates that some examinees in the classes (class 2 and class 4) adjacent to the middle class in the 5-class solution were grouped into the middle class in the 3-class solution. When averaging the means of the two lowest classes in the 5-class solution, the resulting mean was close to the mean for the lowest class in the 3-class solution. Similarly, when averaging the means of the two highest classes in the 5-class solution, the resulting mean was close to the mean for the highest class in the 3-class solution. Obviously, the 5-class solution provided finer distinction between examinees than the 3-class solution.

Five classes were simulated in data generation and the fit indexes generally support the 5-class solution using the mixture Rasch model. The estimated mean and standard deviation for each of the five estimated classes are summarized in Table 4. In general, when the class means were negative, the means were overestimated. On the other hand, when the class means were positive, the means were underestimated. This may indicate a regression towards the mean effect. There is no systematic pattern observed between the true standard deviation and the estimated values.

Based on the estimated means, standard deviations and mixing proportions, the intersecting points for each pair of adjacent distributions can be computed following the details in

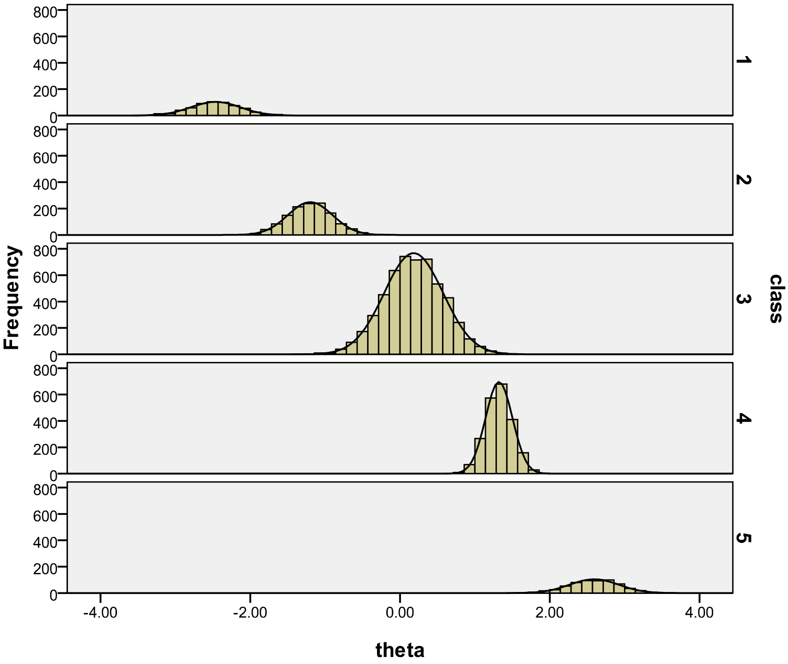
Appendix A. The *x* scores corresponding to these intersecting points were treated as the estimated cut scores for classifying the examinees into five proficiency levels. The true cut scores can be computed in a similar manner. These two sets of cut scores are summarized in Table 5. In general, the estimated intersecting points were higher than the true negative intersecting points and lower than the true positive intersecting points. This again may indicate a regression toward the mean effect.

To better understand the estimated ability and the generated true ability parameters, the generated true ability for each class and the combined classes and the estimated ability parameters based on the Rasch model are plotted in Figures 2 and 3 respectively. Based on the histogram of the generated true ability parameters for each latent class, the five classes separated distinctly from each other although some overlapping did occur. When the distributions were combined and placed on the same latent scale, the clustering around each generated distribution mean was still distinctly different across groups. When the Rasch model was fitted to the data, the estimated ability parameters as shown in Figure 3 did not provide finer information regarding the clustering of the examinees around several theta points along the scale. The reason is that raw score is a sufficient statistics for the Rasch ability estimates. Only 40 items were used in the analysis. The possible number of different estimated theta values is 41. This dramatically constrained the ability estimates which were continuous points along a wider range of theta scale to the limited number of 41 points.

The classifications of examinees within latent classes were made based on a comparison of the estimated ability for each examinee with the theta cut points obtained from the proposed method. The class membership of examinees based on their estimated abilities were compared with their true class membership. The results are summarized in Table 6. The overall classification accuracy is 86.29%. The within-class classification accuracy is 78%, 73%, 92.4%, 83.91%, and 78.17% for class 1 to class 5, respectively. In general, the classification accuracy is relatively high.

Table 5:
True and estimated theta cut scores

Cut Score	True	Estimated
Level 1 vs Level 2	-1.8658	-1.5565
Level 2 vs Level 3	-0.7036	-0.4983
Level 3 vs Level 4	0.9696	0.6944
Level 4 vs Level 5	1.8583	1.5368



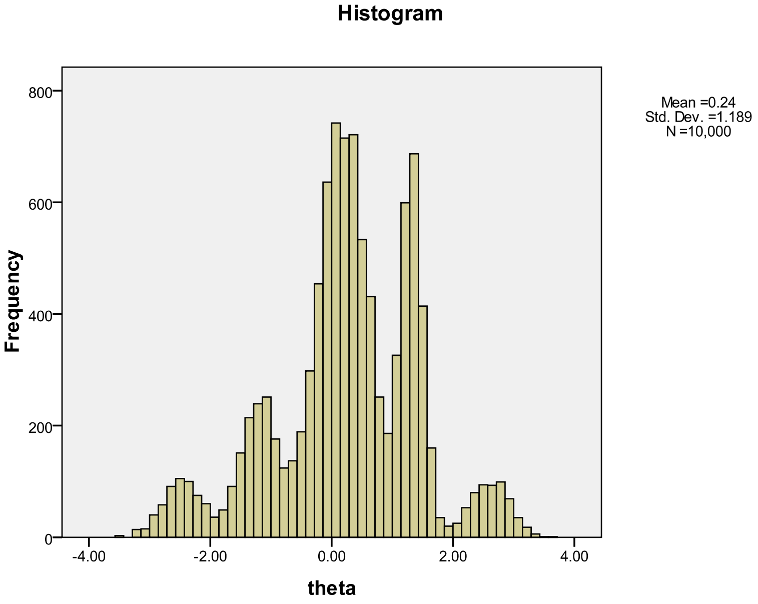


Figure 2:
The distributions of each latent class and the combined distribution for the 5 classes

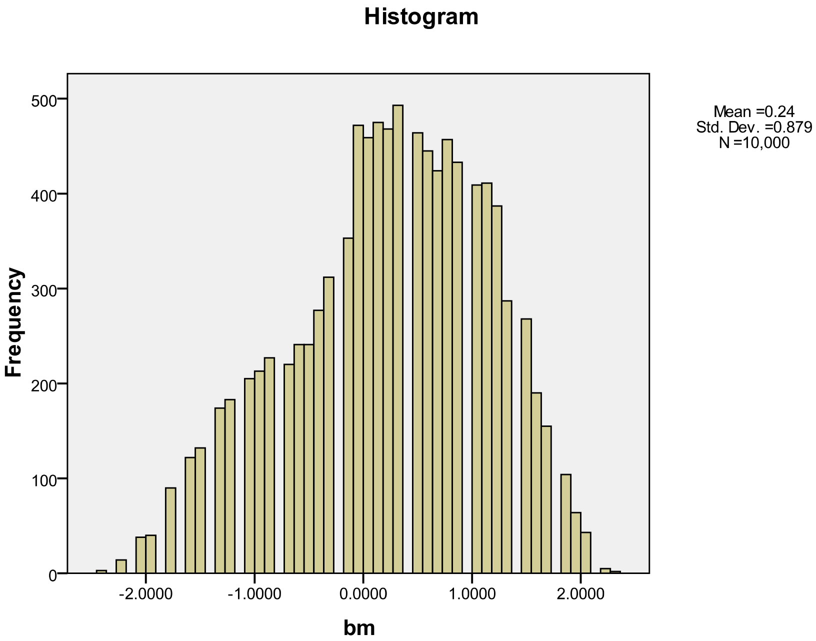


Figure 3:
The distribution of the estimated ability parameters based on the Rasch model

Table 6:
Classification accuracy based on the Rasch model by applying the theta cut scores from the proposed method

			True Class					Total	
			1	2	3	4	5	Count	% of Total
Estimated Class	1	Count	468	134	0	0	0	602	
		% of Total	4.68%	1.34%	.00%	.00%	.00%		6.02%
	2	Count	132	949	135	0	0	1216	
		% of Total	1.32%	9.49%	1.35%	.00%	.00%		12.16%
	3	Count	0	217	4897	230	1	5345	
		% of Total	.00%	2.17%	48.97%	2.30%	.01%		53.45%
	4	Count	0	0	268	1846	130	2244	
		% of Total	.00%	.00%	2.68%	18.46%	1.30%		22.44%
	5	Count	0	0	0	124	469	593	
		% of Total	.00%	.00%	.00%	1.24%	4.69%		5.93%
Total		Count	600	1300	5300	2200	600	10000	
		% of Total	6.00%	13.00%	53.00%	22.00%	6.00%	100.00%	100.00%

Summary and discussions

This paper proposes a method for establishing performance cut scores based on the mixture Rasch model under the assumption that students in different proficiency levels are distinctly different from each other in terms of qualitative characteristics represented by their item response patterns and quantitative characteristics represented by their latent ability along a continuous theta scale. The proposed procedure is based on fitting the mixture Rasch model to the data and using the intersecting point on the corresponding density functions between adjacent distributions to define cut scores for distinguishing between adjacent proficiency levels. Based on one simulation condition intended to reflect a large-scale reading test, the proposed mixture Rasch model based method results in a reasonably high level of classification accuracy.

The advantages of the proposed method for setting performance cut scores are summarized as follow. First, in the current practice of standard setting, there is no statistical validation of the pre-specified numbers of proficiency levels. Policy makers set the number of proficiency levels. Content experts use their knowledge of test content, examinee population, and item information to subjectively come up with the qualitative descriptions of students at different proficiency levels specified by policy makers. There is a lack of a data-driven statistical validation step in the current widely used standard setting methods. The proposed method based on the mixture Rasch model provides empirical evidence related to the validity of the number of proficiency levels set by policy based on both qualitative and quantitative data.

Second, the mixture Rasch model analysis results can help identify the minimally competent or borderline examinees. After finding the number of classes with best fit using the mixture Rasch model with a range of latent classes, the intersecting point between the two adjacent distributions can be established. The corresponding theta values can then be treated as the estimated ability for the minimally competent or borderline examinees in the two adjacent proficiency levels. These examinees can be identified and their performance in the test or item response patterns can be further studied and provide empirical information for fine tuning the cut scores set in the proposed method.

Third, in the process of finding the cut scores using the model based approach, it is easy to cross-validate the cut scores using a subset of items and/or examinees. This cross-validation of cut scores can be a very resource-consuming process in the current practice of standard setting. Further, if proficiency categories are suspected to have changed due to the change in the examinee population, it is relatively easy to re-examine or reset the cut scores based on the proposed method in this study. It is surmised that one requirement will be a large sample of items and examinees to validate the previously set performance standards.

Fourth, the proposed method can be used as a yardstick for comparing multiple standard setting methods irrespective of whether the method is an examinee-centered or test-centered method because the procedure incorporates both test-centered and examinee-centered information. The examinee-centered information is related to the qualitative classification of examinees and the quantitative information of the central location and the variability of each proficiency level and borderline/minimally competent examinees. Test-centered information is related to the item characteristics for examinees in different proficiency categories. The relative magnitudes of item difficulty across latent classes will provide information about which items should be correctly answered by examinees in each proficiency level.

The mixture Rasch model based approach for standard setting distinguishes itself from other model based approaches such as the latent class model (Brown, 2000; Templin et al., 2007), cluster analysis (Sireci, 1995; Sireci et al., 1999), and the attribute hierarchy model (Sadesky & Gushta, 2004) in that other model based approaches still follow a two-step approach. The grouping of examinees is identified in the first stage. Then based on the grouping of examinees, the scores for those examinees at the borderlines or competent and incompetent groups are used to approximate the cut scores. The reason for using the two-stage approach is that the above model based approaches do not estimate the latent ability of examinees. The classification can not be directly related to the theta scale. The proposed approach in this study estimates the group membership of examinees and their latent ability. The latent distributions are scaled on the theta scale. Through the analytical computation, the theta cut points obtained are on the latent ability scale.

There are several limitations associated with this current study that should be delineated. First, in the data generation step, the true mean and standard deviation were simulated by assuming that the intersecting point for two adjacent distributions is two standard deviations away from the mean, which was the midpoint of the two intersecting points. This is a relatively ideal way to simulate test data. Real test data are likely to be less ideally distributed in well separated classes than the simulated study condition here considered. Extensive research reflecting a wider variety of test data characteristics should be further investigated to determine if the procedure here presented tends to work as well under more authentic conditions.

Due to the difficulty in finding a real data set used for standard setting, this study is currently limited to simulation. The proposed method definitely needs to be applied to real standard setting data sets. The results from the proposed method can be compared with those from the current widely used methods for standard setting. The classification consistency can be examined, thus, to gain more insights related to the validity and practicality of the method.

Some other explorations might include whether the performance cut scores obtained from the proposed method can be cross-validated using a subset of examinees or subset of test items. This cross-validation process can be used to evaluate whether the same number of proficiency levels can be identified across multiple samples of items and/or examinees. The implications for the truncated data need further exploration. The effects of sample sizes of examinees and items need to be further investigated. It is speculated that sample sizes for both items and examinees will affect the proper identification of the proficiency levels and the classification of examinees into the correct proficiency category. In particular, when applying the mixture Rasch model, the number correct score is a sufficient statistics for the latent ability estimation. The estimated ability parameters will be constrained to a limited number of values. The impact of sample size may be important and requires further study. Another possible extension is the use of the mixture 2-parameter or 3- parameter IRT models for the proposed mixture IRT model based standard setting procedure. Further, it is possible to extend the procedure to tests consisting only of polytomous items or tests containing both dichotomous and polytomous items. Another relevant issue related to the proposed method is whether this approach should be applied to the item pool or a particular test form conforming to the test specification. It would also be interesting to investigate the method when the test is multidimensional and a mixture multidimensional IRT model will be the calibration model. In addition, the classification decisions based on the proposed method should be compared with those made using other model based methods like latent class analysis (Brown, 2000; Templin et al., 2007) and cluster analysis (Sireci, 1995; Sireci et al., 1999).

Even though the current study uses simulation data, it should be noted that heterogeneity of the examinee population is commonly observed for a test. The data structure we simulated is not atypical in large-scale assessments. Many studies related to mixture IRT models have documented the existence of more then one latent class in real large-scale tests. Cohen et al. (2005) found a three-group solution to the Florida Comprehensive Assessment Test (FCAT; Florida Department of Education, 2002) mathematics test for Grade 9. The three latent classes were different in ability. Li et al. (2009) analyzed grade 3 FCAT math data and a two-class solution in both data sets. The authors analyzed high school graduation test data in Algebra, Biology, and History and found non-normality in the scale score distribution, which is most often an indication of the existence of multiple latent populations. We analyzed each of the three data sets and fitted them with multiple models with different numbers of latent classes. We found 5 latent classes in Algebra; 3 in Biology, and 4 in History.

To successfully implement the proposed method for standard setting, a large sample of items and examinees are recommended to fully capture the possible proficiency categories in the examinee population. Even in later test administrations, examinee samples may vary from time to time, the same proficiency levels can still be legitimately applied. Another reason for using a large sample is that it is likely the mixing proportion may vary from sample to sample. If the mixing proportions are different, it is expected that the intersecting points will be different for two adjacent distributions. Thus, a large sample is recommended to reduce variability. Another issue worthy of attention is related to the statistically significant and distinct classes vs. the practically distinct and different classes. It is reasonable in the real application of the proposed method that some identified latent classes may be statistically significant but not practically important due to the small size of mixing proportion and/or too small mean difference for two adjacent distributions but with very large sample sizes. Careful examinations of the results are needed and their practical implications investigated under these circumstances to evaluate the cut scores from the model based approach.

In summary, this study proposes using the mixture Rasch model to set proficiency cut scores. It can be applied to any test classifying examinees into categories including educational tests, psychological tests, certification/licensure exams, and admission tests. Though further investigation related to the application of the proposed method in real data settings is needed, it is expected that the proposed method can provide model based empirical information related to possible proficiency cut scores. From this, minimally competent borderline examinees can be identified to at least facilitate traditional standard setting and make the judgmental procedure more objective with the empirical data providing initial cut score estimates based on the statistical models. Perhaps such a hybrid method will prove to be the best long-term approach to standard setting.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52,* 317-332.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture model for multiple choice data. *Journal of Educational and Behavioral Statistics*, *26*, 381-409.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331-348.

Boughton, K., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models*. New York: Springer.

Brown, R. S. (2000). *Using latent class analysis to set academic performance standards.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Dai, Y., & Mislevy, R. (2006). *Using structured mixture IRT models to study Differentiating item functioning.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City, NY.

Ebel, R. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Florida Department of Education (2002, March). *Florida Comprehensive Assessment Test.* Tallahassee, FL: Author.

Hambleton, R. K. (1980). Latent ability scores, interpretations, and uses. In S. Mayo (Ed.), *New directions for testing and measurement: interpreting test performance*. San Francisco: Jossey-Bass.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences* (pp. 367-396). Berkeley, CA: McCutchan.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Eds.), *Educational Measurement* (pp. 433-470). Westport, CT: Praeger Publishers.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, *8*, 41-55.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*, 353-366.

Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2010). *Exploring using the Mixture Rasch Model for standard setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1994). *Using actual student work to determine cut scores for proficiency levels: New methods for new tests*. Paper presented at the National Conference on Large-Scale Assessments, Albuquerque, NM.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1995). *Setting standards for performance levels using the student-based constructed-response method*. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*, 307-327.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lewis, D. M., Mitzel, H. C., Green, D. R. (1996). *Standard Setting: A Bookmark Approach*. In D. R. Green (Chair), IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring. Symposium presented at the 1996 Council of Chief State School Officers 1996 National Conference on Large Scale Assessment, Phoenix, AZ.

Li, F., Cohen, A. S., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353-373.

Lissitz, W. L., & Kroopnick, M. H. (2007). *An adaptive procedure for standard setting and a comparison with traditional approaches*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Livingston, S. A., & Zieky, M. J. (1982). *Passing Scores*. Princeton, NJ: Educational Testing Service.

Lu, R., & Jiao, H. (2009). *Detecting DIF using mixture Rasch model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M., & DeChamplain, A. (1998). *Applications of latent class analysis to mastery decisions using complex performance assessments*. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55* (2), 195-215.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Assoc.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide. Fifth Edition.* Los Angeles, CA: Muthén & Muthén.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedogogiske.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Sadesky, G. S., & Gushta, M. M. (2004). *Standard setting using the hierarchy model*. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Sireci, S. G. (1995). *Using cluster analysis to solve the problem of standard setting.* Paper presented at the meeting of the American Psychological Assocation, New York.

Sireci, S. G. (2001). *Standard setting using cluster analysis*. In G. J. Cizek (Ed.), Setting Performance Standards: Concepts, Methods, and Perspectives (pp. 339-354). Mahwah: Lawrence Erlbaum.

Sireci, S. G., Robin, F. R., & Patelis, T. (1999). Using Cluster Analysis to Facilitate Standard Setting. *Applied Measurement in Education, 12* (3), 301-323.

Templin, J., Poggio, A., Irwin, P., & Henson, R. (2007). *Latent Class Model Based Approaches to Standard Setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Taube, K. T. (1997). The incorporation of empirical item difficulty data in the Angoff standard setting procedure. *Evaluation and the Health Professions*, *20*, 479-498.

von Davier, M. (1994). *WINMIRA: A program system for analyses with the Rasch model, with the latent class analysis and with the mixed Rasch model*. Kiel: Institute for Science Education (IPN).

von Davier, M. (2001). *WINMIRA 2001*. Retrieved from http://www.ipn.uni-kiel.de/abt_ppm/tt0506/winmiramanualmvd.pdf

von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: ETS.

Yamamoto, K. (1989). *A Hybrid model of IRT and latent class models. ETS Research Report No. RR-89-41*. Princeton, NJ: Educational Testing Service.

Yamamoto, K. Y., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified HYBRID model. ETS Research Report No. RR-95-16*. Princeton, NJ: Educational Testing Service.

Yamamoto, K. Y., & Everson, H. T. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). Muenster, Germany: Waxmann.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the basic skills assessment tests*. Princeton, NJ: Educational Testing Service.

Appendix A

The function of a normal distribution is

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Assume that 5.9% of the total sample (which is a normal distribution) is located in the lower distribution $N(u_1,\sigma_1)$ and 13.2% of the whole sample is located in the higher distribution $N(u_2,\sigma_2)$. Then we have $N(u_1,\sigma_1)$ with a weight $w_1=0.059$ and $N(u_2,\sigma_2)$ with a weight $w_2=0.132$. Our purpose is to find the intersection potion of the two normal distributions $N(u_1,\sigma_1)$ and $N(u_2,\sigma_2)$. Then we have to solve the following function:

$$w_1\frac{1}{\sqrt{2\pi\sigma_1}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}=w_2\frac{1}{\sqrt{2\pi\sigma_2}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}.$$

That is equivalent to:

$$\frac{w_1}{\sigma_1 e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}}}=\frac{w_2}{\sigma_2 e^{\frac{(x-\mu_2)^2}{2\sigma_2^2}}};$$

which is equivalent to

$$w_2\sigma_1 e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}}=w_1\sigma_2 e^{\frac{(x-\mu_2)^2}{2\sigma_2^2}};$$

which is further equivalent to

$$\frac{w_2\sigma_1}{w_1\sigma_2}=e^{\frac{(x-\mu_2)^2}{2\sigma_2^2}-\frac{(x-\mu_1)^2}{2\sigma_1^2}};$$

Since $a=e^{\ln(a)}$, then $$e^{\ln(\frac{w_2\sigma_1}{w_1\sigma_2})}=e^{\frac{(x-\mu_2)^2}{2\sigma_2^2}-\frac{(x-\mu_1)^2}{2\sigma_1^2}};$$

As a result:

$$\ln(\frac{w_2\sigma_1}{w_1\sigma_2}) = \frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2} \quad (1)$$

This is a typical quadratic equation. For example, if we want to find the intersections of $N(-2.46, 0.33)$ and $N(-1.2, 0.297)$ (with a weight $w_1 = 0.059$ a weight $w_2 = 0.132$.), we can just plug the corresponding means, standardized deviations and weights in to equation (1):

$$\ln(\frac{0.132\times 0.33}{0.059\times 0.297}) = \frac{(x+1.2)^2}{2\times 0.297^2} - \frac{(x+2.46)^2}{2\times 0.33^2}.$$

By solving this quadratic equation: we have

x_1 = -1.876 and x_2 = 10.210

Obviously, the first solution is what we are looking for. Similarly we can find other intersecting points.

Given the generated true distributions, the intersecting points for each pair of adjacent distributions can be computed as follows:

solve $\ln(\frac{0.13\times 0.3296}{0.06\times 0.2974}) = \frac{(x+1.1989)^2}{2\times 0.2974^2} - \frac{(x+2.4625)^2}{2\times 0.3296^2}$;

{x = -1.865816465}, {x = 10.53883360}

solve $\ln(\frac{0.53\times 0.2974}{0.13\times 0.3934}) = \frac{(x-0.1827)^2}{2\times 0.3934^2} - \frac{(x+1.1989)^2}{2\times 0.2947^2}$;

{x = -0.7036014884}, {x = -5.227758315}

solve $\ln(\frac{0.22\times 0.3934}{0.53\times 0.1804}) = \frac{(x-1.3224)^2}{2\times 0.1804^2} - \frac{(x-0.1804)^2}{2\times 0.3934^2}$;

{x = 0.9696004259}, {x = 2.283453720}

solve $\ln(\frac{0.06\times 0.1804}{0.22\times 0.328}) = \frac{(x-2.5943)^2}{2\times 0.3281^2} - \frac{(x-1.3224)^2}{2\times 0.1804^2}$

{x = 1.858348295}, {x = -0.3156921738}

The intersecting points for the estimated distributions:

$$\text{solve } \ln\left(\frac{0.1257\times 0.3123}{0.0626\times 0.3836}\right)=\frac{(x+0.8563)^2}{2\times 0.3836^2}-\frac{(x+2.0368)^2}{2\times 0.3123^2};$$

{x = -1.556522312}, {x = -7.157992419}

$$\text{solve } \ln\left(\frac{0.5286\times 0.3836}{0.1257\times 0.3023}\right)=\frac{(x-0.1228)^2}{2\times 0.3023^2}-\frac{(x+0.8563)^2}{2\times 0.3836^2};$$

{x = -0.4982736026}, {x = 3.955028017}

$$\text{solve } \ln\left(\frac{0.2211\times 0.3023}{0.5286\times 0.1430}\right)=\frac{(x-0.9553)^2}{2\times 0.1430^2}-\frac{(x-0.1228)^2}{2\times 0.3023^2};$$

{x = 0.6943540464}, {x = 1.696226895}

$$\text{solve } \ln\left(\frac{0.0619\times 0.1430}{0.2211\times 0.2419}\right)=\frac{(x-2.4069)^2}{2\times 0.2419^2}-\frac{(x-0.9553)^2}{2\times 0.1430^2};$$

{x = 1.536827443}, {x = -1.186544854}

The intersecting points for the five proficiency levels:

Generated true (weighted)
-1.8658, -0.7036, 0.9696, 1.8583

Estimated (weighted)
-1.5565, -0.4983, 0.6944, 1.5368