A New Method for Standard-setting Using the Rasch Model

Purya Baghaei [*]
Assistant professor of TEFL, English Department, Faculty of Literature & Humanities, Islamic Azad University, Mashhad Branch, Mashhad, Iran
&
Reza Pishghadam [†]
Assistant professor of TEFL, English Department, Faculty of Literature & Humanities, Ferdowsi University of Mashhad, Mashhad, Iran
&
Safoora Navari
MA in TEFL

Abstract
Due to deficiencies of the traditional models of standard setting, this study intends to suggest a new method for setting standards employing Rasch measurement. Precise and efficient methods for setting performance standards and linking tests to ability scales is a much-felt need in today's educational contexts. The introduction of the Common European Framework of Reference as a common paradigm for language teaching and assessment stressed the need for such methods. The suggested method combines the classic test-centered method of standard setting with the probabilistic properties of the Rasch model to set several cut points on the ability continuum. The Wright map which jointly depicts the difficulty location of items and the ability location of persons on a common scale is the cornerstone of this method.

Keywords: Cut-points; Rasch measurement; Standard error of estimates; Standard setting; Wright map

[*] *E-mail address*: puryabaghaei@gmail.com

[†] *E-mail address*: pishghadam@ferdowsi.um.ac.ir

Introduction

In the light of the widespread use of competency testing and due to the decisive role it plays in the interpretation of test results, standard setting has become an increasingly important issue in language assessment. Evidently, the standards are wholly judgmental in nature (Jones and Saville, 2008) and it is by the validity of these judges that the accuracy of the decisions about testees is signified. Due to the inevitability of decision making in all language testes, standards must be set so that by the help of them the levels of performance are clarified for various kinds of decisions (Brown, 2005). Whether the decisions are related to the admission of a student, his placement in an institution, diagnosis purposes, passing a particular level, or getting certification of a course, appropriate cut-points should be set before the test is administered.

Brown (1996) defines standard setting as "the process of deciding where and how to make cut-points" and cut-point as "that score at or above which students will be classified one way and below which they will be classified differently" (p. 249). Based on this definition, setting cut-points seems to be of great importance, because testees above or below the cut-point are classified and treated differently; some testees may be admitted to a specific program and some may not. Evidently, cut-points act like doors; they can open or close windows of opportunities to individuals, shaping their future life. Therefore, setting cut-points is a delicate job which must be handled very carefully.

To set cut-points, testers have used different procedures including: state-centred, teacher-centred, and student-centred methods. These methods have been found faulty in a way or another (Hambleton & Pitoniak, 2006). The authors of this paper believe that Rasch measurement due to its capacity to measure both person ability and item difficulty on a single common scale can be a good method for setting cut-points. Therefore, this study attempts to present a new Rasch-informed model for setting standard procedure which can overcome the old models` shortcomings.

Theoretical Background

Standard setting and Cut-off scores

According to Fulcher and Davidson (2007) cut score is in fact a test score above which a student's *mastery* would be specified or by which the requirements of some *criteria* is clarified. Also it is by the means of level mastery that the *domain score* which is somehow indicative of minimal competence is recognized (Bachman, 1990). By the same token, Hambleton and Eignor (1978) relate the standard setting to the concept of 'competency' and define standard or as they call it 'a minimum proficiency level' as a point on a test score scale that is used to divide testees into two categories, that each of them reflects a distinctive level of proficiency relative to the competency that the test has measured. Also, they label those testees in the higher-scoring category as 'master' or 'competent' and the rest in the lower-scoring category are labelled as 'non-master' or 'incompetent'.

Moreover, they elaborate on "minimum competency testing" as a test which is designed to "determine whether an examinee has reached a prespecified level of performance relative to each competency being measured" and that "pres-specified level or standard may vary from one competency to the next" (p.5). It is derived from such definition that minimal competency tests are considered as a type of competency tests in which standards are introduced to interpret the examinees' performance and that such tests are a special kind of criterion-reference test that are usually utilized in places in need of certification.

Methods for Establishing Standards

In order to make appropriate decisions which deal with the lives of students, testers have to utilize the techniques and methods for setting standards which have been promising and well-established. As it was needed in the educational testing field, some assorted methods have been proposed which all share judgmental bedrock. The selection of each method depends on the merits and limitations of different methods in different contexts. Brown (1996) elaborates on three types of standard setting methods which are used in educational testing. These

types include 'state mastery', 'test-centered' and 'student-centered' methods.

State-mastery Methods

The first sort of methods considers the mastery of the students in a trait being measured as the blueprint for passing that exam. These 'state mastery' methods are subject to some crucial criticism. Meskauskas (1976) for instance observed that state mastery methods make a sharp distinction between the group of students who have mastered the material 100% and those who have not mastered it at all (Brown, 1996). Actually, by viewing students as purely white or black, such methods seem problematic in language teaching and thus in decision making areas.

Test-centered Methods

Considering one of the major drawbacks of 'mastery methods', that is their attempt to dichotomize learning into 100% mastery or 0% mastery on the grounds of a set of scores, Brown (1996) asserted that the need for a number of 'continuum methods' which would not neglect the continuous nature of test scores is felt. So since the continuous nature of scores were ignored in the last group of methods, a number of 'continuum methods' came into existence (Brown, 1996). This set of methods like the rest is based on judgments, with the only distinction of focusing on the 'content' of the test rather than the student performance. In order to clarify the difference between the state and continuum methods, Meskauskas (1976) describes the two significant characteristics of continuum methods with an underlying basis of 'ability'. He holds that in continuum methods mastery is considered as an ability or set of abilities which are continuously distributed and that at the upper end of the continuum there exists an area and an individual is termed as a master when he has the ability to equal or exceed the lower bound of this area (Hambleton and Eignor, 1978).

According to Hambleton and Pitoniak (2006) in test-centered methods judges decide on the level of performance which is required to

meet each performance standard. This is done by making judgments about the expected performance on each item for hypothetical examinees that just barely fulfils the requirements for a certain performance standard (Nasstrom and Nystrom, 2008). Brown (1996) points to four key test-centered methods for standard setting as Nedelsky, Anghoff, Ebel, and Jeager methods.

The first test-centered method, proposed by Nedlesky (1954), relying on judgments of test design fell short of being appropriate for tests other than multiple-choice ones. In this method the judges were responsible for viewing the items in the test with a criterion in mind and that criterion includes the response option that the minimally competent student would be able to omit as incorrect (Hambleton and Eignor, 1978). In 1971 Anghoff suggested another test-centered method which had the estimation of the probability of the level of performance of a competent student as its core. To apply the Anghoff method to tests with items scored as right or wrong, the judges are asked to conceptualize a group of barely qualified testees and after that to estimate the proportion of this group which would answer each item in the test correctly (Cizek, 2006). Then, for each judge the estimated probabilities are summed and those sums are averaged across judges to arrive at a recommended cut-score (Ferdous and Plake, 2007).

The use of Angoff method to represent the test-centred methods is in fact so outstanding among the researchers that along with its original version, the modified and the extended version have been introduced and have elevated it to the most widely used procedure for standard-setting (Hurtz and Auerbach, 2003). Furthermore, although the Angoff method was originally conceived as a one-stage test-centered process, MacCann and Stanley (2006) believe that this method has now typically developed into a multi-stage procedure where the judges make independent judgments and after that discuss their initial decisions.

In fact, they hold that there exists a natural affinity between IRT models and many standard-setting procedures like Angoff. They claim

there is a shared view of a continuum of achievement and a probabilistic definition of mastery on an item in these methods. They also report that Both van der Linden (1982) and Kane (1987) have discussed the similarities they consider to exist between IRT models and Angoff standard-setting procedures (cited in MacCann and Stanley, 2006).

To name the advantages of the Angoff method we can claim that it is easy to administer, that it yields compensatory cut-scores '(i.e. a high score on one item can balance a low score on another item (Hambleton and Pitoniak, 2006)', and also the fact that the method can be implemented before the administration of the test (Kane, 1998). Although beside the above mentioned characteristics, this method has the privilege of being useful for tests with other than multiple-choice items, there are still some drawbacks which lower its applicability.

First, according to Brown (1996) it lacks the possibility of being utilized for items that are not scored dichotomously. Second, there is the difficulty for the judges to estimate the performance on individual items for a group of just barely qualified testees, and finally as Hambleton and Pitoniak (2006) mentioned there exists the tendency to overestimate performance on less difficult items and underestimate difficult items by judges (Nasstrom and Nystrom, 2008).

Ebel's (1979) method makes a judgment about the success of test items based on their *relevance* (questionable, acceptable, important, essential) by *difficulty* (easy, medium, hard). In this method, judges classify the items into a two-way taxonomy of difficulty and relevance, after that each judge gets to decide on the ability of a minimally qualified respondent to answer correctly to a particular percentage of items in a cell. On the next level, the number of items in each cell is multiplied by the respective percentage, and the results are summed across the 12cells (Goodwin, 1996). Actually, the cut-off score would be the average of these figures for all judges. Although Ebel's method enjoys the advantage of involving some judges in working on a common scale; in applying this method there occurs some problems, for instance

the judges would face difficulty in keeping the two dimensions of the taxonomy separately in their mind, also it seems somehow time-consuming (Brown, 1996).

And finally in this group, Jeager's method (1982) going through an iterative process to develop a consensus is considered to be more elaborate than the previous ones. In addition to using judges from a variety of backgrounds, this method employs normative data. Actually, rather than asking questions involving 'minimal competence' that are believed to be hard to operationalize and conceptualize, Hambleton and Eignor (1978) report Jeager's questions as:

> "Should every high school graduate be able to answer this item correctly?" "----Yes, ----No." and "If a student does not answer this item correctly, should he be denied a high school diploma? "----Yes, ----No." (p. 28)

Thus, in this method after judges undergo a series of iterative processes, some normative data are presented and then the standards which are set by all groups of judges are pooled. On the next step, a median is computed for each judge. Finally, the minimum median across all groups is chosen as the standard. In contrast to the complexity, this method has the advantage of asking multi-interested group of judges to take part in the decision making process.

Student-centered Methods

Another category of continuum methods which relies on the students' performance is of two types: borderline-group and contrasting-group methods. Zieky and Livingston (1977) proposed a method whose judgments clarify the borderline cases of students in a given population which would specify a borderline performance in that category. In the borderline-group method, judges are asked to conceptualize the characteristics of border-line examinees and identify specific examinees that fit these characteristics. Then according to Cizek (2006) the assessment is administered, after scoring and analyzing, the medium

score of those who are defined as borderline examinees are typically used as the cut-score (Nasstrom and Nystrom, 2008). Here, unlike the former methods, the observation of teachers is given credit and becomes the basis for establishing the cut-points.

Due to the advantages of borderline-group method, Jeager (1989); Hambleton and Pitoniak (2006) point to its conceptual simplicity and Kane (1998) recommends its usage for holistic and constructed-response tests. Along with the problems of taking much time (Kane, 1998), requiring large panel of judges (Hambleton and Pitoniak, 2006) and a large number of testees (Cizek, 2006), Livingstone and Zieky (1989) indicate a potential problem with the borderline-group method that the cut-score arrived at by teachers with high-performing testees usually tends to be higher than the cut-scores from teachers with lower-performing classes (Nasstrom and Nystrom, 2008).

Moreover, Hambleton and Eignor (1978) hold that in this method the weak point of guessing the minimally competent student by the judges is circumvented and that can add to the reasons that keep these methods in favor.

The last group of student-centered methods suggested by Zieky and Livingston (1977) is contrasting-group method which draws on the construct validity strategies and works with establishing cut-points based on the performance of acceptable and inadequate category of students. In fact, Hambleton and Eignor (1978) hold that the Border-line group and Contrasting-group methods are procedurally similar and that they just differ in the sample of students on which the performance data is collected. In this method after judges have defined minimally acceptable performance for the particular subject area, they are required to identify those students they believe to be definite masters or non-masters of the test skills. Evidently, 100 students are suggested by Zieky and Livingston for the smaller group in order to gain more stable results.

On the next step, the test score distributions for the two groups are plotted and then the point of intersection is decided to be the initial standard. In order to reduce what they call "false masters", that is, the students who are identified as masters by the test, but who have not adequately mastered the objectives or "false non-masters", the students who are identified as non-masters by the test, but who have adequately mastered the objectives, Zieky and Livingston (1977) propose to adjust the standard by relying on the relative seriousness of the two types of errors. This method seems more satisfactory due to its connection to intervention and differential-group strategies of construct validity which makes it closer to the purpose of the test itself (Brown, 1996). Moreover, Buckendahl (2006) believes that although this method can be utilized independently, it is also possible that this method be used as a complement to the informed judgment or other standard setting methods.

Empirical Background

Hambleton and Eignor (1978) reported some of the studies which have been conducted in the field of testing that require the use of standard setting. Livingston (1975) among others has presented a procedure which is based on linear or semi-linear utility functions. In his study he pinpoints the use of these utility functions in viewing the effects of decision-making accuracy with the basis of a particular performance standard. In another study, Livingston (1976) introduced a method that was used for selecting standards by stochastic approximation techniques and was dependent on the standards setting for that measurement. Huynh (1976) uses an external criterion as a basis for standard setting method for a competency test.

Some of the other studies attempted to make comparisons between the standard setting methods by stating their merits over each other in different situations. For instance, Andrew and Hecht (1976) carried out a comparison of the Nedlesky and Ebel methods by setting standards for 180 items in two separate occasions using those different methods and then comparing the results. Also, more recently Näsström and Nyström (2008), with respect to procedural, internal and external evidence, have

evaluated and compared a version of Angoff method as a representative of the class-centered method and the borderline-group method which represents a standard setting method with examinee-centered procedure.

Also, considering the importance of an appropriate method for standard setting on a given situation, Jeager (1976); Zieky and Livingston (1977) and Popham (1978) have provided some guidelines for selecting such methods based on the merits of different methods (cited in Hambleton, Powell, & Eignor, 1979). Elsewhere, Bejar (2008) attempts to illustrate the importance of standard setting with reference to accountability testing in K-12 and proposes that some questions that have emerged concerning standard setting in a k-12 context can be addressed by considering standard setting as an integral aspect of the test development process, the point that has not been standard practice in the past.

Hambleton (1999) by the use of performance categories for score reporting has presented detailed steps for setting performance standards on educational assessments. Furthermore, due to important role the current proficiency tests play in the future of the examinees and the widespread use of such tests around the world, some researchers have tried to discuss the standard setting methods employed by test developers. Papajohn (2006), among others, with concern to the speaking component, has dealt with some standard setting processes for the Next Generation TOEFL Academic Speaking Test (TAST).

Rasch Measurement

Rasch Merits

Rasch model as a branch of IRT, is considerably useful in 'rater-mediated testing environments' and also it provides room for 'investigation of rater characteristics and task characteristics on scores'. Rasch model, which is often called as a 'latent trait' model, has this underlying assumption that views each test continuum as a *latent trait* upon which learners, items and also criterion levels of ability that is *standards* can be located (Jones & Saville, 2008). One of features of

Rasch model is that in this model the discrimination of all the items are supposed to be equal and that there exists no guessing. In addition to TOEFL test; most of the IRT applications to language testing have so far used the Rasch model (Bachman, 1990).

The Rasch model is the simplest of the IRT models, considering that it uses only one parameter to describe an item which is its difficulty. MacCann and Stanley (2006) go on to say that the one-parameter logistic Rasch model as a way of standard setting is different from more complex IRT models in that such models lack a unique one-to-one relationship between total score and ability. In addition, it is held that under Rasch modelling, there exists a *line of relationship* between ability and total score, actually for a given ability there is one associated total score while in non Rasch models, for a given ability there is a distribution of total scores, and vice versa.

The use of Rasch analysis can yield opportunities for spotting the inconsistent individual rater behaviour and that Rasch model can consider all the factors that would in a way or other affect the student's final score such as his ability, the severity of the rater and also the task difficulty. In fact, in a test of checking the abilities of the examinees, it is crucial that the performance examination be able to measure the candidates' abilities consistently. The efficiency of this matter, according to Lunz and Wright (1997) can be best improved by using a latent trait model, Rasch model which provides the examinee with an ability estimate that is independent of the present value of the examinee facet elements, that is the judges, tasks and items (Weir, 2005).

Rasch Model Details

The Rasch model is a probabilistic measurement model which states: the probability that person *n* gets item *i* right is a function of the difficulty of item *i* and the ability of person *n*. This probability is governed by the difference between the ability of person *n* and the difficulty of item *i*.

$$P_{ni}(X=1) = f(\theta_n - \delta_i)$$

Where P_{ni} (X=1) is the probability that person n gets item i right

f is function

θ_n is the ability of person n

δ_i is the difficulty of item i

If $\theta_n - \delta_i > 0$ then $P_{ni}(X_i=1)>0.50$

That is, if the person's ability is greater than the item' difficulty the probability of getting the item right for this person is greater than 50%.

If $\theta_n - \delta_i < 0$ then $P_{ni}(X_i=1) <0.50$

That is, if the person's ability is below item's difficulty the probability of getting the item right for this person is smaller than 50%.

And if $\theta_n - \delta_i = 0$ then $P_{ni}(X_i=1) =0.50$

That is, if the person's ability is equal to the item's difficulty the probability of getting the item right for this person is exactly 50%.

The exact value of the probability that a person with a known ability estimate gets an item with a known difficulty estimate right can be calculated by means of the simple logistic function which is written as:

$P_{ni}(X_i=1|\ \theta_n, \delta_i) = e^{(\theta_n - \delta_i)}/[1+e^{(\theta_n - \delta_i)}]$

This is read: the probability of person n getting item i right, given the person's ability θ_n and the item's difficulty δ_i is equal to e (the base of natural logarithm, equal to 2.73) raised to the power of ability minus difficulty divide by same value plus one. This equation is called the simple logistic function. This function is cumulative, i.e., as the ability increases relative to difficulty the probability of a correct reply increases. Person ability estimates, item difficulty estimates, the standard errors of these estimates and fit statistics are all derived from this equation after some performing some complicated mathematical operations.

Now, suppose that a person with ability 5 attempts an item with difficulty 3. Here the difference between ability and difficulty is 2, which is greater than 0. So we can expect a probability greater than 50%

for this person to solve this item. Inserting the numbers in the equation we will have:

$P = e^{5-3} / [1 + e^{5-3}] = 0.88$

If a person with ability 2 attempts an item with difficulty 3, the probability that s/he gets this item right is:

$P = e^{2-3} / [1 + e^{2-3}] = 0.27$

And if a person with ability 5 attempts an item with difficulty 5 we will have:

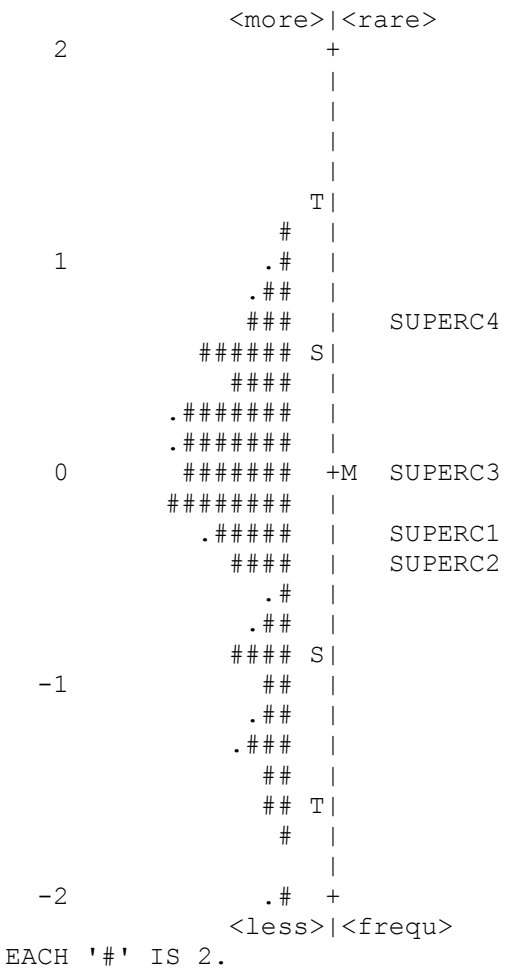
$P = e^{5-5} / [1 + e^{5-5}] = 1/2 = 0.50$

Computation of the probability that person *n*- with a known ability estimate- gets item *i*- with a known difficulty estimate- right, implies that the ability and difficulty estimates are both expressed on a common scale. In fact, this is one of the interesting properties of the Rasch and all IRT models that the ability of persons and the difficulty of items can directly be compared (Baghaei, 2009).

Wright Map

Most Rasch software graphically depict item difficulty and person ability estimates on a graph referred to as *item-person map* or more recently the *Wright map* (Wilson, 2005), named after the famous Rasch proponent Benjamin D. Wright. Figure 1 below is an example Wright map. The codes on the right represent the items and the '#' on the left represent persons. The calibrated vertical line in the middle is the variable of interest or the construct we want to measure. "M" represents the origin or the zero point of the scale, which is set by default at the mean difficulty of the items. S's are placed one standard deviation above and below the mean and T's are placed two standard deviations above and below the mean.

Figure 1
Wright map produced by Winsteps Rasch program for 4 items and 150 persons



By looking at the map one can locate the ability of persons and the difficulty of items by reading off the scale values on the vertical line.

More difficult items and more proficient persons are located on the upper part of the scale. Less difficult items and less proficient persons fall at the lower parts of the scale.

The Wright map can provide interesting information about the test. Narrow spread of items indicates poor coverage of the test construct. It is important to note that the entire set of the items that comprise the test might have a wide spread, however, there might be some gaps in some regions of the scale, which indicate poor coverage of the construct in that region. This signals that the persons who fall at that part of the scale are not measured precisely and the test developer needs to add more items to cover the empty or scantily covered parts of the scale. The map can also show whether the test is well-targeted for the sample. The more the test difficulty is geared to the ability level of the persons, the smaller the errors of estimates and the more precise the measurement will be. If the bulk of items line up with the bulk of persons, then the test is well targeted for the sample. If the majority of the items cluster on the top of the scale and the majority of persons at the bottom or vice versa then the test is too difficult or too hard for the respondents (Wright, 1997).

The Proposed Method

The probabilistic properties of the Rasch model when combined with test-centred method of standard setting along with the visual facilities that the Wright map provides can offer great help in setting cut-off scores to indicate different levels of performance. The method suggested here requires judgments about the minimum ability level required of hypothetical examinees to answer each item correctly. Needless to say, this minimum ability is expressed in terms of the levels or the scale to which we want to link our test. Therefore, the following procedure is suggested to set cut points:

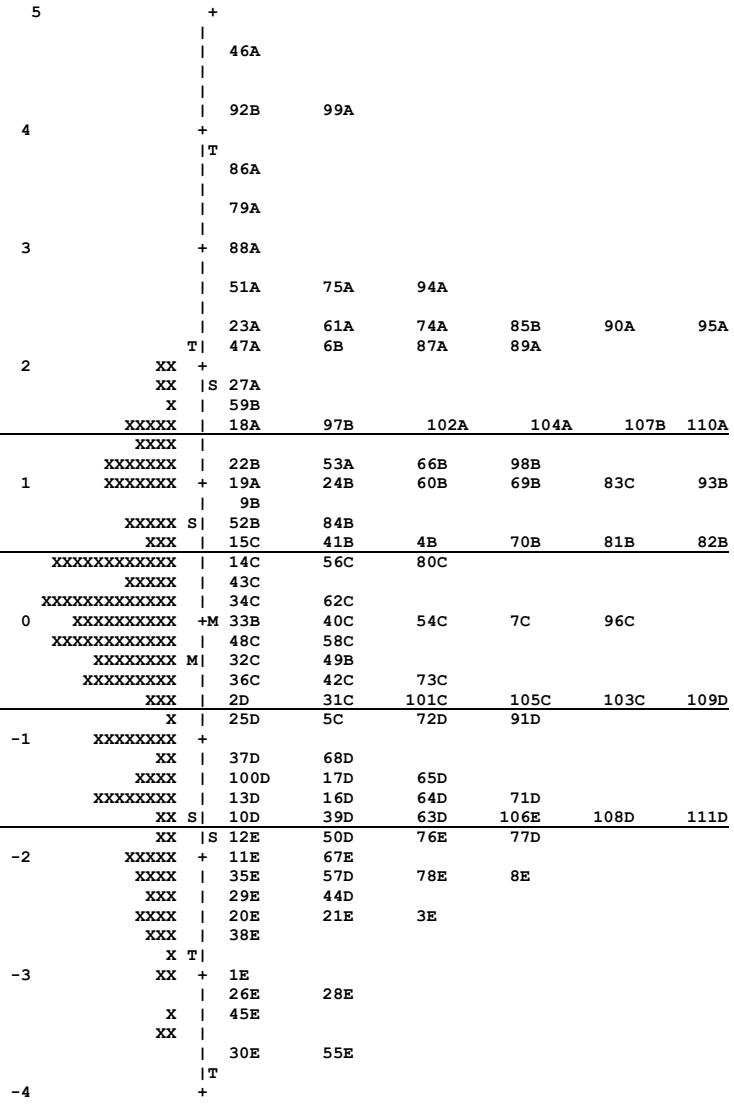
1. Judges assign items to the levels of ability. In other words, they decide what ability level a person should be to answer each item. If there are say, five levels of proficiency from A to E, A being the highest and E the lowest, then the items are rated on a five point scale from 1-5. Five

corresponding to the lowest level, E and 1 to the highest level, A. For instance, if the judges agree that "border line" Level B students can answer an item correctly the item is rated 2 and if they envisage that for answering the item a test-taker should be at least a border line Level A student then the item is rated 1. The average judge ratings for an item is considered as its final difficulty estimate. All the items are rated in this way and assigned to one of the levels on the scale.

2. The items are administered to a group of test-takers and Rasch analysed to obtain their difficulty estimates. The success and preciseness of the standard setting procedure heavily depends on the accordance between the judge-envisaged item difficulties and empirical student-based item difficulties. Any standard setting procedures in which this accordance is not achieved is futile. If the judges have done their job properly then there must be a correspondence between the empirical item estimates and judge-based item difficulties.

Figure 2 shows the item estimates hierarchy on an item-person map.

Figure 2
Difficulty order of items and their judge-based corresponding levels



The Level A items are clustered at the top of the map and the other levels' items are ordered accordingly. However, there are some items which are misplaced. It is obvious that judge-intended levels of the items never correspond exactly with the Rasch measures. For instance, as can be seen in Figure 1, Rasch has reported some A items down in the B region (or below) and some B items up in the A region (or above).

Standard-setting always requires a compromise between the judges' item hierarchy and the empirical (Rasch) item hierarchy which corresponds to actual examinee performance. Standard-setting also requires negotiation about the location of the criterion levels. There will be several reasonable positions for the criterion level, from least-demanding to most demanding.

We might choose the transition points to be the lines on which the minimum number of items is misclassified between two adjacent levels. For example, the transition point between Level A and Level B is the point where the items predominantly become Level B items (as is done in Figure 1). That is, the difficulty level of item 18A or 97B which is 1.53 logits. As stated before, person ability estimates and item difficulty estimates are expressed on a common scale. Persons whose ability measure is equal to the difficulty of items 18A and 97B have 50% chances of getting these item right. This is a legitimate reason to consider 1.53 logits as the transition point for Level A.

3. One can also be more stringent and choose the ability level required to have 60% chances of success on the items in the transition points to be the cut-off score. As we saw, the items at the transition points between Level A and Level B have a difficulty estimate of 1.53 logits. This is an item of borderline difficulty. In other words, an ability estimate of 1.53 logits can be the minimum cut-off score for Level A. This is the ability level required to have 50% chances of getting this item right. To be on a safe side, one can also define:
"cut-off score" = 60% chances of success on an item of borderline difficulty

Therefore, the cut-off score for Level A will be:

$P_{ni}(X_i=1|\ \theta_n,\ \delta_i)= exp(\theta_n - \delta_i)/[1+ exp(\theta_n - \delta_i)]$

$0.60= e\exp(\theta_n -1.53)/[1+ e\exp(\theta_n -1.53)]$

$\lg 1.5=\theta_n -1.53$

$\theta_n =1.93$

4. The cut-off scores for the other levels can be determined in a similar way. The items at the point of transition between Level B and Level C are 15C, 41B, 4B, 70B, 81B, 82B with difficulty estimates of 0.68 logits. Therefore the cut-off score for Level B can either be 0.68 logits, if we consider the 50% chances of success on the items at the transition point as the minimum requirement to be a Level B examinee, or 1.08 logits if we consider 60% chances of success on the items at the transition point as the minimum requirement for this level.

Conclusion

As it was stated, there exist multiple ways for standard setting which are problematic in a way or another. It seems that these methods are not exact in determining the cut points. Therefore, the authors in this study have introduced a novel and straightforward method for linking tests to ability scales. The proposed method employs the theoretical and graphical features of the Rasch model to both set cut-offs and check the accuracy of the link. The Wright map serves as the cornerstone of the suggested method.

The empirical and judge-based ordering of items on an interval scale is used to set performance standards. When the two orderings match to an acceptable degree, the transitions are the points where judge-based levels change from one to another. These points have location calibrations that can be read off the scale and adopted as cut-offs.

The proposed model seems to enjoy some advantages. First, due to the probabilistic nature of Rasch measurement, it can locate cut points with more accuracy than the traditional methods. Unlike the classical

test theory in which there is only one standard error of measurement for the entire sample of the examinees, the Rasch model uses a unique standard error of measurement which is associated with every item and person estimate. These standard errors of estimate can help to know the precision of an estimate. The precision of a cut-point is the same as the precision of a person measure at the cut-point. If that person measure standard error is too large, then more items at the level of the cut-point are needed.

Second, it can be used to check the accuracy of the procedure. If the transition-point items have poor fit, then this threatens the construct-validity of the instrument, and so the reliability (precision) and accuracy (did the correct people pass or fail?) of the pass-fail decision. When calibrating items for standard setting, persons who are clearly misbehaving (guessing, carelessness, etc.) should be omitted from the analysis.

Third, the Rasch cut-points can also be translated into their raw score equivalents, a feature that does not exist in two and three parameter IRT models. Since examinees with identical raw scores can have different ability estimates under these IRT models.

Thus, due to the aforementioned merits of the proposed method, it is recommended as a model for those who like to come up with right and exact decisions in categorizing individuals into more or less competent in a specific domain of knowledge.

Acknowledgements
We gratefully acknowledge the project reported here was supported by a grant-in-aid of research from the Iran `s National Institute of Elites to the second author without which this research would not have been possible.

Received 5 March, 2009
Accepted 23 June, 2009

References

Andrew, B. J. & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards, *Educational and psychological measurement,* 36, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Eds.), *Educational measurement* (2nd ed.). Washington, D.C.: American council on education.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Baghaei, P. (2009). *Understanding the Rasch mode*l. Mashad: Mashad Islamic Azad University Press.

Bejar, I. I. (2008). Standard Setting: What Is It? Why Is It Important? *Educational testing service,* 7, 20-26.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum.

Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall Regents.

Brown, J. D. (2005). *Testing in language programs* (2nd ed.). New York: McGraw-Hill.

Buckendahl, C. D. (2006). Adaptation within a language: Considerations for standard setting. *Paper presented at the International Test Commission conference, Brussels, Belgium.*

Cizek, G. J. (2006). Standard Setting. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development.* Mahwah: Lawrence Erlbaum Associations.

Ebel, R. L. (1979). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Ferdous, A. A. & Plake, B. S. (2007). Item selection strategy for reducing the number of items rated in an Angoff standard setting study, *Educational and Psychological Measurement,* 67, 193-206.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment*. London: Routledge.

Goodwin, L. D. (1996). Focus on quantitative methods: Determining cut-off score, *Research in Nursing & Health,* 19*,* 249-256.

Jeager, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application, *Educational Evaluation and Policy Analysis,* 4*,* 461-476.

Jones, N. & Saville, N. (2008). Scales and Frameworks. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of educational linguistics* (pp. 495-509). UK: Blackwell.

Hambleton, R. K. (1999). Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.1-32). Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K. & Eignor, D. L. (1978, October). Competency Test Development, Validation, and Standard-Setting. *Paper presented at the Minimum Competency Testing Conference of the American Education Research Association, Washington, DC.*

Hambleton, R. K., Powell, S., & Eignor, D. L. (1979). Issues and methods for standard setting, *Laboratory of Psychometric and Evaluative Research Report*, 70, 30-108.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments, *Applied Psychological Measurement,* 24, 355-366.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Eds.), *Educational Measurement* (4th ed.). Westport: American Council on Education & Praeger Publishers.

Hurtz, G. M. & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus, *Educational and Psychological Measurement,* 63, 584-601.

Huynh, H. (1976). Statistical consideration of mastery score, *Psychometrika,* 41*,* 56-78.

Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedures, *Journal of Educational Measurement*, 24*,* 333-345.

Kane, M. (1998a). Choosing between examinee-centred and test-centred standard-setting methods, *Educational Assessment,* 5*,* 129-145.

Linacre, J. M. (2007) *A user's guide to WINSTEPS-MINISTEP: Rasch model computer programs*. Chicago, IL: winsteps.com.

Lunz, M. E. & Wright, B. D. (1997). Latent trait models for performance examinations. In J. Rost & R. Langeheine (Eds.), *Applications of latent and latent class models in the social science.* Retrieved from : *http://www.ipn.unikiel.de/aktell/htm.*

MacCann, R. G. & Stanley, G. (2006). The Use of Rasch Modeling to Improve Standard Setting, *Journal of practical assessment, Research and Evaluation*, 11*,* 1-17.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting, *Review of Educational Research,* 45, 133-158.

Näsström, G. & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items, *Journal of practical assessment, Research and Evaluation,* 13*,* 1-12.

Nedlesky, L. (1954). Absolute grading standards for objective tests, *Educational and Psychological Measurement,* 14*,* 3-19.

Papajohn, D. (2006). Standard setting for next generation TOEFL Academic Speaking Test (TAST): Reflections on the ETS Panel of International Teaching Assistant Developers, *TESL-EJ,* 10*,* 35-40.

Popham, W. I. (1978). *Setting performance standards.* Los Angeles: Instructional objectives exchange.

Rasch, G. (1960/1980). *Probabilistic models for intelligence and attainment tests*. University of Chicago Press, Chicago, Illinois.

Rasch, G. (1966). An item analysis which takes individual differences into account, *British Journal of Mathematical and Statistical Psychology*, 19, 49-57.

van der Linden, W. J. & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer.

Weir, C. J. (2005). *Language testing and validation*. New York: Macmillan.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wright, B. D. (1997). Fundamental measurement for outcome evaluation. MESA Research Memorandum, No. 66. Retrieved from: *www.rasch.org/memo66.htm.*

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