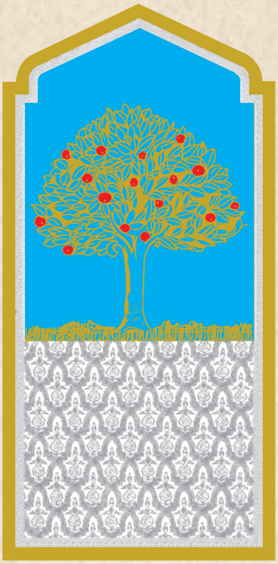


AL-SHAJARAH

JOURNAL OF ISLAMIC THOUGHT AND CIVILIZATION
OF
THE INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA (IIUM)

SPECIAL ISSUE:
EDUCATION

2018

AL-SHAJARAH

Al-Shajarah is a refereed international journal that publishes original scholarly articles in the area of Islamic thought, Islamic civilization, Islamic science, and Malay world issues. The journal is especially interested in studies that elaborate scientific and epistemological problems encountered by Muslims in the present age, scholarly works that provide fresh and insightful Islamic responses to the intellectual and cultural challenges of the modern world. *Al-Shajarah* will also consider articles written on various religions, schools of thought, ideologies and subjects that can contribute towards the formulation of an Islamic philosophy of science. Critical studies of translation of major works of major writers of the past and present. Original works on the subjects of Islamic architecture and art are welcomed. Book reviews and notes are also accepted.

The journal is published twice a year, June-July and November-December. Manuscripts and all correspondence should be sent to the Editor-in-Chief, *Al-Shajarah*, F4 Building, Research and Publication Unit, International Institute of Islamic Civilisation and Malay World (ISTAC), International Islamic University Malaysia (IIUM), No. 24, Persiaran Tuanku Syed Sirajuddin, Taman Duta, 50480 Kuala Lumpur, Malaysia. All enquiries on publications may also be e-mailed to alshajarah@iium.edu.my. For subscriptions, please address all queries to the postal or email address above.

Contributions: Submissions must be at least 5,500 words long. All submissions must be in English or Malay and be original work which has not been published elsewhere in any form (abridged or otherwise). In matters of style, *Al-Shajarah* uses the *University of Chicago Manual of Style* and follows the transliteration system shown on the inside back cover of the journal. The Editor-in-Chief reserves the right to return accepted manuscripts to the author for stylistic changes. Manuscripts must be submitted to the Editor-in-Chief in Microsoft Word. The font must be Times New Roman and its size 12. IIUM retains copyright to all published materials, but contributors may republish their articles elsewhere with due acknowledgement to *Al-Shajarah*.

ISSN 1394-6870

USING THE MANY-FACET RASCH MODEL TO DETERMINE CUTSCORES AND RESOLVE FUNDAMENTAL STANDARD SETTING ISSUES[1]

Noor Lide Abu Kassim
Kamal J. I. Badrasawi
Nor Zatul-Iffa

Abstract

This paper illustrates the use of the Many-facet Rasch model and the Objective Standard Setting Method in the construction of cutscores, and in dealing with fundamental standard setting issues. With the proliferation of standardized testing for accountability purposes, the tracking of educational growth within and across nations, and the high-stakes use of educational standards, issues pertaining to the selection of the "right" standard setting method for the construction of defensible cutscores has become more and more prominent. The Many-facet Rasch Model used together with the Objective Standard Setting Method allow for the modeling and adjustment of rater severity in the standard setting process, and facilitate clear and efficient identification of inconsistent judges and misjudged items. However, as with most model-based approaches, a clear understanding of construct theory is imperative to arrive at valid and defensible cutscores and performance standards.

Keywords: Cutscore, Many-Facet Rasch Model, English language test, Objective Standard setting Method

[1] Article received: October 2018; Article submitted: November 2018; Article accepted: December 2018

Introduction

Performance standard, which deals with the question of "how good is good enough" with respect to the attainment of educational standards, has been the subject of considerable attention, and considered to be one of the most controversial issues in educational measurement. If excessively high standards are set, failure to attain the set standards could result in unwarranted sanctions for schools[2] as well as inequitable penalties on students. Conversely, if excessively low standards are set, detrimental consequences on the value of education will result.[3] Despite the controversy that shrouds the use of performance standards, there are legitimate grounds for their use in educational decision-making[4] in contexts where assessments are used for certification or licensure, performance standards are deemed essential.[5] What is considered a minimal level of competency needs to be clearly ascertained to "protect the public from incompetent practitioners".[6] Though problems of misclassifications cannot be

[2] Robert Linn and National Center for Research on Evaluation, Standards and Student Testing, "Performance Standards: Utility for Different Uses of Assessments," *Education Policy Analysis Archives* 11, no. 31 (2003), accessed October 6, 2018, https://epaa.asu.edu/ojs/article/download/259/385.

[3] James Popham, *Modern Educational Measurement: Practical Guidelines for Educational Leaders* (LA: Allyn & Bacon, 2000).

[4] Ronald Hambleton, "Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process," in *Setting Performance Standards: Concepts, Methods, and Perspectives,* ed. Gergory Cizek (Mahwah, NJ: Lawrence Erlbaum, 2001): 89-116; Popham, *Modern Educational Measurement:* . . .; Gregory Cizek, "Conjectures on the Rise and Fall of Standard setting: An Introduction to Context and Practice," in *Setting Performance Standards: Concepts, Methods, and Perspectives,* ed. Gregory Cizek (Mahwah, NJ: Lawrence Erlbaum, 2001): 3-18; Linn and National Center for Research on Evaluation, Standards and Student Testing, "Performance Standards: Utility for . . ."

[5] Linn and National Center for Research on Evaluation, Standards and Student Testing, "Performance Standards: Utility for . . ."

[6] Ibid., 2

avoided,[7] standards still need to be set "[as] there are legitimate practical reasons that require that a decision be made".[8]

Performance standards are also essential to provide invaluable feedback for continued curricular and instructional improvement.[9] They allow for "tracking progress of achievement for schools, states or the nation"[10] and more importantly, for the monitoring and improvement of student learning. In the classroom context, performance standards provide educators with a diagnosis of what is lacking and the corrective measures that need to be taken as a result of acceptable or unacceptable performance.[11] The setting of performance standards inevitably involves human judgment and, therefore, is not infallible. However, this does not mean that the setting of educational standards should be avoided as standards are crucial in the educational decision-making process. What needs to be borne in mind is that there must be clear and valid reasons for the use of performance standards in order to avoid undesirable consequences.

With greater demands for higher quality education and accountability in student learning and achievement, considerable efforts have been made towards ensuring the process of setting educational standards, a rational and defensible one. This is evidenced by the voluminous literature on standard setting methods[12]

[7] Ebel, as cited in Robert Hambleton, "On the Use of Cut-off Scores with Criterion-Referenced Tests in Instructional Settings," *Journal of Educational Measurement* 15, no.4 (1978): 277-290.

[8] Linn and National Center for Research on Evaluation, Standards and Student Testing, "Performance Standards: Utility for . . ., ", 2.

[9] Nancy Burton, "Performance Standards," (A paper prepared under a grant from Carnegie Corporation of New York, 1997), *Journal of Multi Disciplinary Evaluation* 7, no. 15 (2010): 159-170, accessed October 12, 2018. http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/301; Robert Linn, "Assessment and Accountability," *Educational Researcher* 29, no. 2 (2000): 4-16.

[10] Linn and National Center for Research on Evaluation, Standards and Student Testing, "Performance Standards: Utility for . . .,", 3.

[11] Burton, "Performance Standards . . .," 159-170.

[12] Ronald Berk, "A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests," *Review of Educational Research,* 56, no. 1 (1986): 137-172; R. M. Jaeger, "Certification of Student Competence," in *The American Council on Education/Macmillan Series on Higher Education. Educational Measurement*, ed. R. Linn (New York, NY, England: Macmillan Publishing Co, Inc;

and the validation of set standards.[13] However to date the issue of the "right" standard setting method has remained unresolved and may never be resolved as it has been clearly established that different standard setting methods yield different results.[14]

With the increased use of constructed response (CR) items and performance assessment in high-stakes standardized testing; the need for multiple cutscores to differentiate differing levels of achievement; the mounting concerns around fairness and legal issues; and the emphasis on meeting rigorous cutscores; the selection of the "right" standard setting method has been further complicated. In response to these new demands, existing methods have been modified and 'more defensible' methods developed. For example, methods that were initially used for selected-response (SR) items, such as the Angoff procedure, have been modified[15] to include the judgment of CR

American Council on Education, 1989), 485-514; Benjamin Wright, "How to Set Standards," accessed October 12, 2018, https://www.rasch.org/memo77.pdf; Gregory Cizek, "Conjectures on the Rise and Fall of Standard setting: . . .", 3-18; Elizabeth Manias and Tim McNamara, "Standard Setting in Specific-Purpose Language Testing: What can a Qualitative Study add?" *Language Testing* 33, no.2 (2015): 235-249; Thomas Eckes, "Setting cut Scores on an EFL Placement Test using the Prototype Group Method: A Receiver Operating Characteristic (ROC) Analysis," *Language Testing* 34, no.3 (2016): 383-411.

[13] Michael Kane, "Validating the Performance Standards Associated with Passing Scores," *Review of Educational Research* 64, no.3 (1994); Michael Kane, "So Much Remains the Same: Conception and Status of Validation in Setting Standards," in *Setting Performance Standards: Concepts, Methods, and Perspectives,* ed. Gergory Cizek (Mahwah, NJ: Lawrence Erlbaum, 2001):53-88; Gregory Stone, "Objective Standard Setting," (PhD dissertation, Chicago, IL: The University of Chicago, 1996); Jone Norcini and Judy Shea, "The Credibility and Comparability of Standards," *Applied Measurement in Education* 10, no.1 (1997): 39-59; Ronald Hambleton, "Setting Performance Standards on Educational Assessments . . ." 89-116; National Research Council, *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress* (National Academies Press, 1999), accessed October 13, 2018, https://bit.ly/2yDA2tI.

[14] Jaeger, "Certification of Student Competence . . .", 485-514; Michael Kane "So Much Remains the Same: Conception and Status of Validation in Setting Standards," in *Setting Performance Standards: Concepts, Methods, and Perspectives,* ed. Gergory Cizek (Mahwah, NJ: Lawrence Erlbaum, 2001): 53-88; Cizek, Gregory, "Conjectures on the Rise and Fall of Standard setting: . . .", 3-18.

[15] Yoko Kozaki, "Using GENOVA and FACETS to set Multiple Standards on Performance Assessment for Certification in Medical Translation from Japanese into

items. Newly developed methods include the Bookmark Procedure, the Body of Work Method (BoW), the Analytic Judgment Method, Cluster Analysis, the Integrated Holistic Judgment Method, the Item-mapping Method and the Objective Standard Setting Method.[16]

From these new developments, one important observation can be made: the prevalent use of "model-based" approaches in recent years.[17] The increasing use of model-based approaches can be attributed to the robustness of these approaches in dealing with both SR and CR type items, performance assessments, as well as other advantages over the Classical Test Theory in relation to measurement issues. Nonetheless, it is important to highlight that in most of these approaches, the full potential of the measurement model in resolving assessment and standard setting issues has not been fully addressed. This paper puts forward a standard setting procedure that makes use of the full potential of the Many-facet Rasch model in the construction of cutscores and in dealing with fundamental issues in standard setting.

Fundamental Issues in Standard Setting

In selecting the "right" standard setting method, several issues are of primary concern. The first relates to the judgment task that judges or panelists are required to perform. The *Standards for Educational and Psychological Testing* (AERA, APA, NCME) has made a very clear stand on this issue.[18]

English," *Language Testing Journal* 21, no. 1 (2004): 1-27; H.C. Mitzel et al., "The Bookmark Procedure: Psychological perspectives," in *Setting Performance Standards: Concepts, methods, and perspectives,* ed. Gergory Cizek (Mahwan, NJ: Lawrence Erlbaum, 2001): 249-282.

[16] Cizek, Gregory, "Conjectures on the Rise and Fall of Standard setting: . . .", 3-18.

[17] Suzan Loomis and Mary Bourque, "From Tradition to Innovation: Standard Setting on the National Assessment of Educational Progress," in *Setting Performance Standards: Concepts, Methods, and Perspectives,* ed. Gergory Cizek (Mahwan, NJ: Lawrence Erlbaum, 2001): 89-116.

[18] American Educational Research Association, American Psychological Association & National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, D.C.: American Psychological Association, 1999), 60.

> When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

In "non-objective methods"[19] such as Nedelsky's and Angoff's, the judgment task requires judges or panelists to estimate the probability that a minimally competent examinee will succeed on test items. This is considered ineffectual as judges are asked to perform a task that "is too difficult and confusing" and nearly cognitively impossible.[20] A more serious criticism of this judgment task is that it draws the focus of judgment from the content, and therefore the measured construct, to the prediction of examinee performance on test items.[21] Methods such as these begin with content, but "ends up atomized into hundreds of contentless score fractions devoid of a clear and meaningful description of the standard."[22]

A second issue in determining the utility of a standard setting method is its capacity in dealing with diverse item types.[23] Constructed response items are fast becoming a common feature in most high-stakes assessment programmes. One of the criticisms that have been leveled at some widely-used standard setting methods is that these methods are only relevant for use with particular item types; namely, SR items (e.g., the Nedelsky method). It is, therefore, important to examine the generalizability of the standard setting method to different item types or formats. The capacity of the standard setting method to handle a combination of item types is yet another important concern.[24] This is particularly so when

[19] Stone, "Objective Standard Setting . . .", 187-201.

[20] National Research Council, *Grading the Nation's Report Card: . . .*; National Centre for Education Statistics, accessed February 24, 2004, https://nces.ed.gov/.

[21] Stone, "Objective Standard Setting . . ."

[22] Gregory Stone, "Standard Setting Methods," *Rasch Measurement Transactions* 9, no. 3 (1995), accessed October 13, 2018, https://www.rasch.org/rmt/rmt93m.htm.

[23] Mitzel et al., "The Bookmark Procedure: Psychological perspectives . . ." 249-282.

[24] Ibid.

performances on different subtests consisting of different item types (e.g. SR and CR items) are to be combined and used in a compensatory manner. When different subtests are used to assess competency in separate domains or subconstructs, the usual practice is to set cutscores or criterion levels on each individual subtest. This conjunctive approach is used when examinees are expected to meet the requirement of each domain.[25] When different item types are necessary to assess competency in the different domains, the use of the conjunctive approach is not problematic. However, this is not so when decisions are required with the use of multiple subtests consisting of different kinds of content and item type.

The next pertains to variability in judgment. It is a major issue in standard setting as judges represent different stakeholders with diverse background and expectations. In the standard setting process, there are three main sources of judgment-related variability. The first source of variability relates to inconsistencies between judges' prediction of examinee probability of success and the empirical item difficulty estimate (in the case of methods such as Angoff's). This type of judgment variability is typically corrected by introducing performance data as well as iterations in the judging process. The second source of variability relates to judges' internal consistency in making judgment. A review of newly-developed and long-standing standard setting methods indicates no clear strategies or procedures for the identification of this unwanted variability in judgment.

The third source of variability involves the stochastic nature of judge rating.[26] As interrater agreement is usually desired, iterations and panel discussions are introduced in most standard setting methods to facilitate reconciliation of rating differences and increase interjudge agreement. Another approach to deal with this variability is to adjust the final cutscores using the standard error from the judgment process.[27] Though independent expert opinion is encouraged and differences in judges' background characteristics are desired to arrive at the most representative and acceptable cutscores,

[25] Beret Green, "Setting Performance Standards," (presentation, MAPAC Meeting, 2000).

[26] Mike Linacre, *Many-Facet Rasch Measurement* (Chicago: MESA Press, 1989).

[27] Stone, "Objective Standard Setting . . ."

disagreement in judgment is often looked upon as error.[28] This is unproductive as the “fundamental principle of the expert panel is that it seeks to bring together individuals with a wide array of experiences to produce the most credible, most generalizable results” (p.5). In Rasch measurement, this type of variability is considered inevitable as it is a natural outcome of judge independency.[29] And as this notion of judge agreement is virtually impossible to achieve in practice, it would be much more productive to instead focus and improve on internal judge consistency.[30]

To meet the present demands of complex assessment programmes, the rising concerns pertaining to fairness in assessment, and the emphasis on the robustness of cutscores, many new standard setting methods have been developed. However, these newly developed methods have not satisfactorily addressed the issue of judge variability in standard setting. This study illustrates how the Many-facet Rasch Model together with the Objective Standard Setting method can be used not only for the construction of cutscores but also to deal with core fundamental issues in standard setting.

From the Islamic perspective, Muslims are urged to make accurate decisions by adopting the virtue of *Adl* (justice), which is considered as one of the core or noble virtues,[31] mooted in the Noble Qur’an and Hadiths. Allah (S.W.T) says: “Allah commands justice, the doing of good, and liberality to kith and kin, and He forbids all shameful deeds, and injustice and rebellion: He instructs you, that ye may receive admonition” (Qur’an 16:90)[32]. Allah (S.W.T) also says “Allah doth command you to render back your Trusts to those to whom they are due; And when ye judge between man and man, that ye judge with justice: Verily how excellent is the teaching which He giveth you! For Allah is He Who heareth and seeth all things” (Qur’an 4:58). The following Hadith clearly urges and motivates people to have justice. It was narrated from ‘Abdullah bin

[28] Ibid.

[29] Linacre, *Many-Facet Rasch Measurement* . . .

[30] Ibid.; Stone, “Objective Standard Setting . . .”

[31] Shams al-Din Sarkhasi, *al-Mabsut* 14 (59-60) .

[32] The English translation for all the Qur’anic verses is taken from *The Meanings of The Holy Qur’an*, trans Abdullah Yusuf Ali, accessed October 13, 2018, http://www.islam101.com/quran/yusufAli/.

‘Amr bin Al-‘As that: The Prophet (S.A.W) said: “Those who are just and fair will be with Allah, Most High, on thrones of light, at the right hand of the Most Merciful, those who are just in their rulings and in their dealings with their families and those of whom they are in charge.”[33]

‘Adl (Justice) has various related words and each one has its meaning or aspect of justice as *qist* (equity), *istiqmah* (correctness), *insaf* (fairness) *mizan* (balance/scale) etc.[34] However, the most frequently used term is ‘*adl,* and it includes fairness and equality.[35] ‘*Adl* (justice) refers to the act of being just and fair through putting things in their rightful or correct place and giving others equal treatment.[36] Abdullah and Nadvi further explains that Allah (S.W.T) has not prescribed specific ways or means, but general guidelines, on how to establish justice. Importantly, Qaradawi stresses that all the means, procedures, and methods used to accomplish justice must be valid (i.e. in line with Islamic law).[37]

For example, the Prophet Shu’aib instructed his people to be just and use fair measurement/scale because they cheated each other and showed injustice in their business transactions. Allah (S.W.T) says: “To the Madyan People (We sent) Shu’aib, one of their own brethren: he said: O my people! worship Allah. Ye have no other god but Him. And give not short measure or weight: I see you in prosperity, but I fear for you the penalty of a day that will compass (you) all round. And O my people! give just measure and weight, nor withhold from the people the things that are their due: commit not evil in the land with intent to do mischief.” (Qur’an 11: 83-84).

Methodology

Participants: The standard setting panel consisted of 14 judges who were language instructors as well as item writers. The judges

[33] *Sunan an-Nasa’i* 5379 Book 49, Hadith 1, (English Translation, 6, Book 49, Hadith 5381), accessed October 10, 2018, https://sunnah.com.

[34] Majeed Khadduri, *The Concept of Justice*, 1st edn. (in Arabic) (Dar Al Hasad: Damascus, 1998).

[35] Ibid.

[36] Mohammad Abdullah and Mohammad Nadvi, “Understanding the Principles . . .”

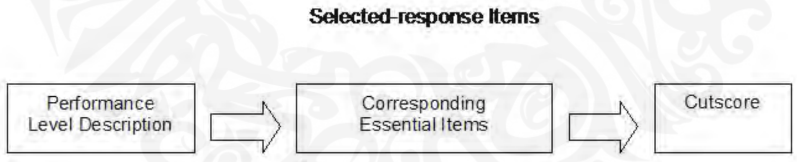
[37] Ibid., 276.

possessed at least a basic degree in TESOL or a master's degree in a similar field, and at least 2 years teaching experience. Out of the 14 judges, two could not complete the judgment process and hence, their ratings were not included in the analysis.

Instrument/Tests: The instrument used in this study was a placement battery developed for purposes of student exemption from and placement into four English language support courses. The placement battery consisted of 3 subtests: Paper 1 (Grammar & Reading), Paper 2 (Essay Writing) and Paper 3 (Speaking). In this study, the focus was on the first subtest. Paper 1 had 75 multiple-choice questions, 40 of which were grammar items and 35 were reading comprehension items.

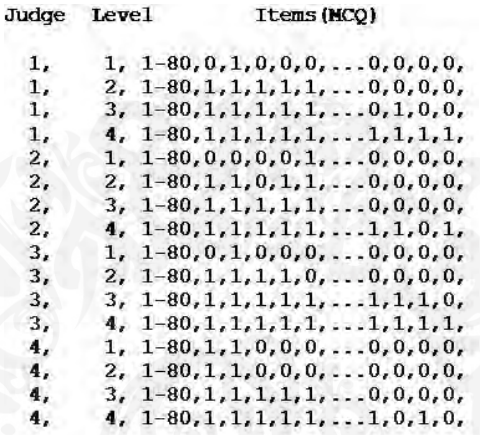
Procedures: The Objective Standard Setting Method (Stone, 1996) was used to generate judges' ratings. For the multiple-choice subtest (Paper 1), judges were asked to individually select essential items for each of the four cutscores. Essentiality of items is referenced against verbal descriptions of what a minimally competent student is expected to be able to do/know in order to be classified as having achieved a given cutscore (or standard). Items that were selected as essential were marked as 1 and non-essential items were marked as 0 (Figure 1).

Figure 1: Translation of Performance Level Description into Corresponding Cutscores for SR Items



Judges' identification number, target cutscore (level) and item ratings were put into a data matrix consisting of three facets: Judges (judges/panel involved in the standard setting study), Level (the specified cutscore/ criterion level of performance), and Items (items on the test). At the onset of the standard setting process, judges were trained on how to interpret the item specifications that were used in the writing of the grammar and reading subtests. Figure 2 presents part of the resulting data matrix used in the Facets analysis.

Figure 2: Data Matrix for the Facets Analysis Quantification of the qualitative or evaluative decisions:



```
Judge  Level        Items(MCQ)

 1,     1, 1-80,0,1,0,0,0,...0,0,0,0,
 1,     2, 1-80,1,1,1,1,1,...0,0,0,0,
 1,     3, 1-80,1,1,1,1,1,...0,1,0,0,
 1,     4, 1-80,1,1,1,1,1,...1,1,1,1,
 2,     1, 1-80,0,0,0,0,1,...0,0,0,0,
 2,     2, 1-80,1,1,0,1,1,...0,0,0,0,
 2,     3, 1-80,1,1,1,1,1,...0,0,0,0,
 2,     4, 1-80,1,1,1,1,1,...1,1,0,1,
 3,     1, 1-80,0,1,0,0,0,...0,0,0,0,
 3,     2, 1-80,1,1,1,1,0,...0,0,0,0,
 3,     3, 1-80,1,1,1,1,1,...1,1,1,0,
 3,     4, 1-80,1,1,1,1,1,...1,1,1,1,
 4,     1, 1-80,1,1,0,0,0,...0,0,0,0,
 4,     2, 1-80,1,1,0,0,0,...0,0,0,0,
 4,     3, 1-80,1,1,1,1,1,...0,0,0,0,
 4,     4, 1-80,1,1,1,1,1,...1,0,1,0,
```

The two-facet Rasch model given in Engelhard and Stone (1998) for the evaluation of the quality ratings given by standard setting judges is as follows:

$$Pr\{x_{ni}=1 \mid \beta_n, \delta_i\} = \exp(\beta_n - \delta_i) \,/\, [1 + \exp(\beta_n - \delta_i)].$$

The above model can also be expressed as follows:

$$\log(P_{ni1}/P_{ni0}) = \beta_n - \delta_i$$

where:

P_{ni1} = probability of judge *n* giving a rating of 1(essential) on item *i*

P_{ni0} = probability of judge *n* giving a rating of 0 (not essential) on item *i*

β_n = view of specialists in the field from judge *n*.

δ_i = judged essentiality of item i.

The simple general form of MFRM (Linacre, 1989) is given as:

$$\log \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = B_n - D_i - C_j - F_k$$

Where:

P_{nijk} is the probability of examinee *n* being awarded on item *i* by judge *j* a rating of *k*

P_{nijk-1} is the probability of examinee *n* being awarded on item *i* by judge *j* a rating of *k-1*

B_n is the ability of examinee *n*

D_i is the difficulty of item

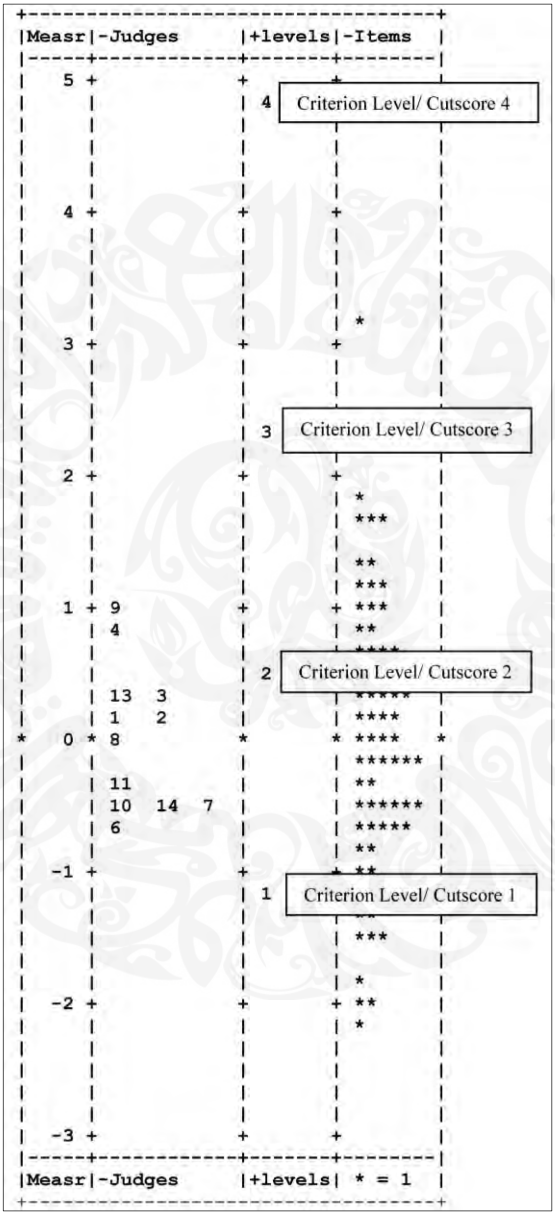
C_j is the severity of judge *j*

F_k is the extra difficulty overcome in being observed at the level of category *k*, relative to category *k-1*.

Results

Facets calibrations of judges, criterion levels (cutscores) and items are presented in Figure 3. The first column is the logit scale followed by the standard setting judges, levels (cutscores), test items and the scale used for rating of essay performances. From judges' distribution in Figure 3, it is evident that there is some variation in judges' perception of essential items. The separation index (3.01) and the chi-square value of 119.6 with 11 df, significant at p <.01 indicate that judges consistently differ from one another in overall severity of judgment (Table 1). Judge 9 is seen to be the most severe and judges 10 and 6 the least severe. With the exception of judge 9, all the judges cluster within -0.5 to +0.2 logit. Figure 3 also indicates that the four criterion levels are clearly separated. However, the criterion levels for 3 and 4 are exceptionally high, in relation to the distribution of test items. The first cutscore appears to be low; however, it is still above some of the easiest items on the test. How the lowest cutscore relates to the actual examinee distribution will be examined later in this paper.

Figure 3: Distribution of Calibrated Items, Judges and Initial Cutscores



```
+------------------------------------+
|Measr|-Judges     |+levels|-Items   |
|-----+------------+-------+---------|
|   5 +            +       +         |
|     |            | 4  Criterion Level/ Cutscore 4
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|   4 +            +       +         |
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|     |            |       | *       |
|   3 +            +       +         |
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|     |            | 3  Criterion Level/ Cutscore 3
|     |            |       |         |
|   2 +            +       +         |
|     |            |       | *       |
|     |            |       | ***     |
|     |            |       |         |
|     |            |       | **      |
|     |            |       | ***     |
|   1 + 9          +       + ***     |
|     | 4          |       | **      |
|     |            |       | ****    |
|     |            | 2  Criterion Level/ Cutscore 2
|     | 13  3      |       | *****   |
|     | 1   2      |       | ****    |
*   0 * 8          *       * ****    *
|     |            |       | ******  |
|     | 11         |       | **      |
|     | 10  14  7  |       | ******  |
|     | 6          |       | *****   |
|     |            |       | **      |
|  -1 +            +       + **      |
|     |            | 1  Criterion Level/ Cutscore 1
|     |            |       | **      |
|     |            |       | ***     |
|     |            |       |         |
|     |            |       | *       |
|  -2 +            +       + **      |
|     |            |       | *       |
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|     |            |       |         |
|  -3 +            +       +         |
|-----+------------+-------+---------|
|Measr|-Judges     |+levels| * = 1   |
+------------------------------------+
```

Table 1, on the other hand, gives a detailed judge measurement report. Although the difference in judge severity is quite small (about 1.3 logits from the most severe to the least severe), there is a significant variation in judge severity. In terms of judges' self-consistency, Judges 2, 4 and 8 appear to be clearly misfitting.

Table 1: Judge Measurement Report

Judge	**Observed Score (Count)**	**Measure (S.E.)**	**Infit MNSQ**	**Outfit MNSQ**
Judge 1	196 (296)	0.17 (0.17)	1.10	0.90
Judge 2	199 (300)	0.17 (0.17)	1.14	**1.61**
Judge 3	195 (300)	0.28 (0.17)	1.24	1.22
Judge 4	164 (284)	0.87 (0.17)	1.36	**2.43**
Judge 5	Not Included			
Judge 6	219 (284)	-0.74 (0.18)	0.99	0.73
Judge 7	219 (292)	-0.54 (0.15)	1.09	0.77
Judge 8	208 (300)	-0.08 (0.17)	1.18	**1.81**
Judge 9	168 (299)	0.98 (0.16)	0.90	0.73
Judge10	225 (300)	-0.58 (0.17)	1.19	1.11
Judge 11	216 (300)	-0.31 (0.17)	1.17	1.07
Judge 12	Not Included			
Judge 13	194 (300)	0.30 (0.15)	1.13	1.12
Judge 14	223 (300)	-0.52 (0.15)	1.08	0.98

Separation: 3.01; Reliability: .90

Fixed (all same) chi-square 119.6 d.f.: 11 Significance: .00

The criterion levels (cutscores) measurement report (Table 2) indicates a significant difference between levels. The separation index is 13.81, chi-square value is 917.0 with 3 df, significant at *p*<.01. The criterion level (cutscore) to separate candidates into the first and second level of English language performance is calibrated at -1.14 logits whereas the highest cutscore that exempts examinees from the language support courses is calibrated at 4.90 logits.

Table 2: Criterion Levels (Cutscores) Measurement Report

Criterion Levels	**Measure (S.E.)**	**Infit MNSQ**	**Outfit MNSQ**
Level 4	4.90 logits (0.29)	1.03	1.09
Level 3	2.30 logits (0.11)	1.13	1.34
Level 2	0.44 logit (0.08)	1.11	1.24
Level 1	-1.14 logits (0.08)	1.18	1.16

As regards item displacement, five items (Items 39, 63, 69, 73, 27, 67, 75, and 30) showed positive displacement values of above 2. Item 39 showed the highest positive displacement estimate indicating that judges had misjudged this item, which is easy (measure: -2.13 logits) as being a difficult item and hence assigned it to a high level of language performance. Item 5 (measure: +1.12 logits), which has a displacement estimate of -3.12 logits, on the other hand, has been misjudged as an easy item. Overall, there are more items that have been misjudged as difficult (as indicated by the empirical calibrations) than those that have been misjudged as easy. With respect to fit, items that showed high displacement values also had high Infit and Outfit Mean Square estimates (Table 3). This displacements are evidence that there are considerable inconsistencies in judges' estimation of the difficulty levels of the items.

Table 3: Item Fit Statistics and Displacement Values

Item	Measure (S.E.)	Infit MNSQ	Outfit MNSQ	Displacement
5	1.12 (0.41)	2.21	2.64	-3.12
46	0.49 (0.42)	1.58	1.26	-2.47
22	0.22 (0.41)	1.35	1.03	-2.20
25	0.23 (0.41)	1.58	1.23	-2.20
21	-0.31 (0.41)	1.13	0.70	-2.08
39	-2.13 (0.55)	4.43	4.62	5.23
63	-1.33 (0.46)	4.25	9.00	5.18
69	-0.02 (0.41)	2.23	3.69	3.13
73	-0.18 (0.41)	2.28	7.59	3.04
27	-1.75 (0.50)	2.57	2.56	3.02
67	0.50 (0.40)	1.68	3.36	2.43
75	0.93 (0.44)	1.62	1.70	2.38
30	-1.22 (0.45)	1.64	1.42	2.00

Figure 4 displays the estimated cutscores (criterion levels) derived from the Facets analysis as applied to actual examinee and item distributions. It is evident that the judges' had underestimated the lowest cutscore (Level 1), and overestimated the highest cutscore (Level 4). The underestimation of the lowest cutscore or criterion level could be attributed to the substantial number of easy items (off-target items) on the test while the overestimation of the highest cutscore (Level/Cutscore 4) could be due to judges' high expectation of examinee performance on the test.

To test this supposition another analysis was carried out using judges' ratings of only the grammar items. Figure 4 presents judges' ratings of essential items for the grammar test items. From the figure, it is apparent that the standard setting judges had selected a small number of items to represent the first cutscore. About half of the items were selected to represent the second cutscore. As for the third cutscore more than 90% of the items were selected. For the fourth cutscore almost all the items were selected as essential for students to have knowledge of, in order to be considered competent enough to be exempted from the language support courses.

Figure 4: Judges' Rating of Essential Items by Cutscore (level)

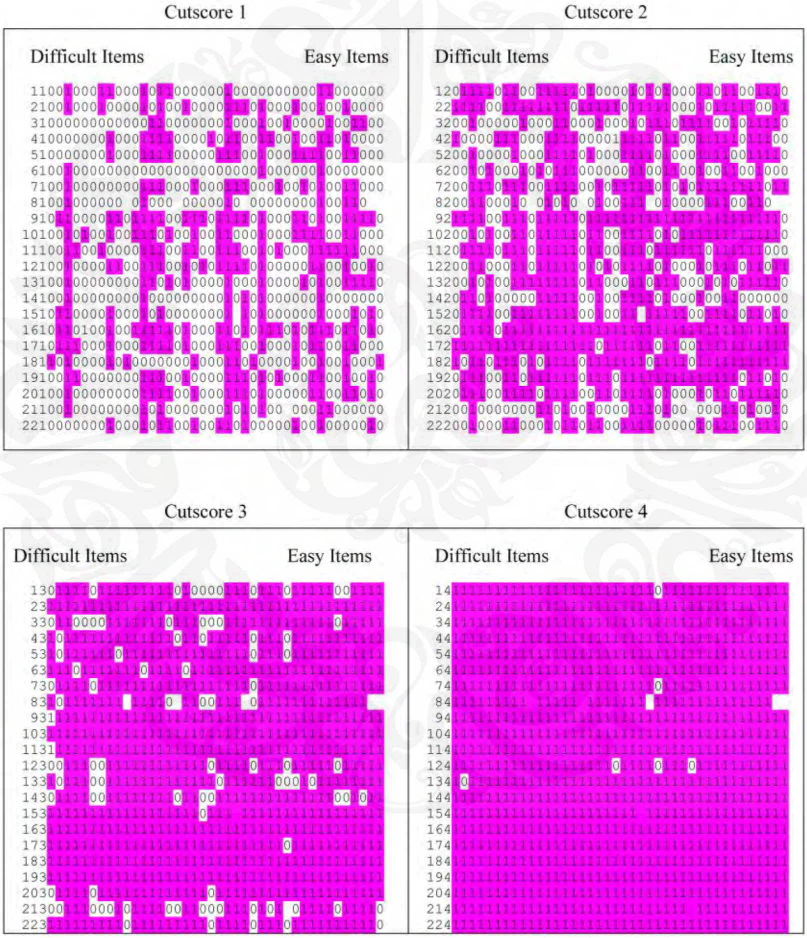
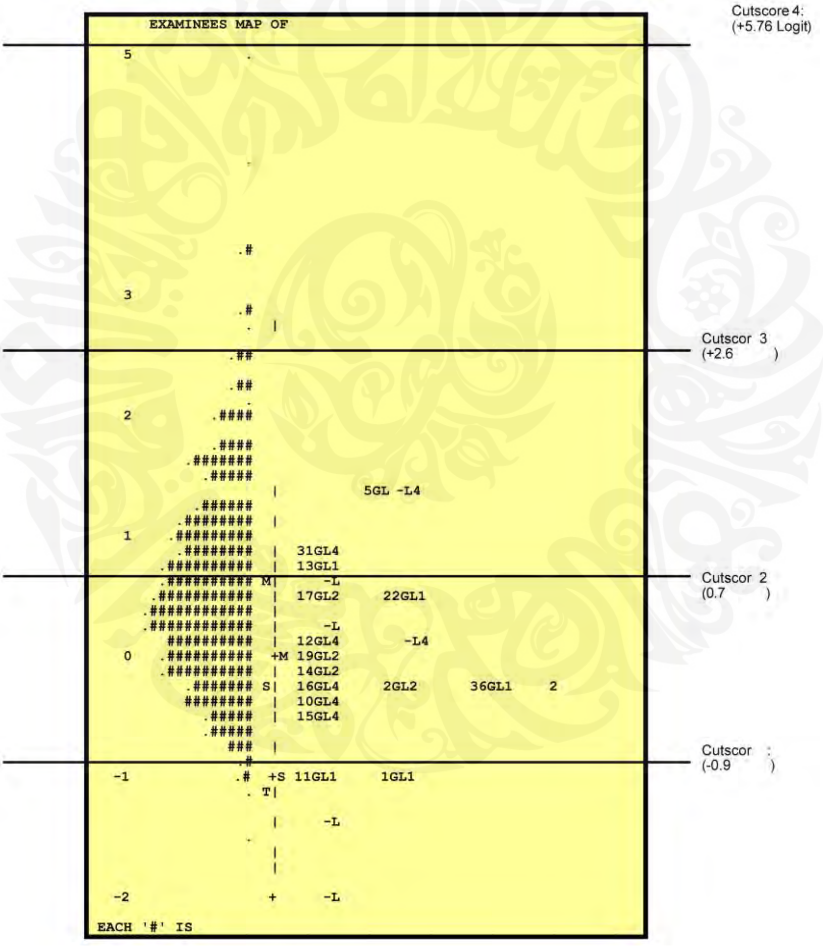


Figure 5 displays the resulting cutscores as applied to the examinee and item distributions generated from an analysis of examinee responses to the grammar test items. The underestimation of the first cutscore and the overestimation of the third and fourth cutscores are obvious. Given judges' selection of items for the four cutscores, this pattern of cutscores is expected.

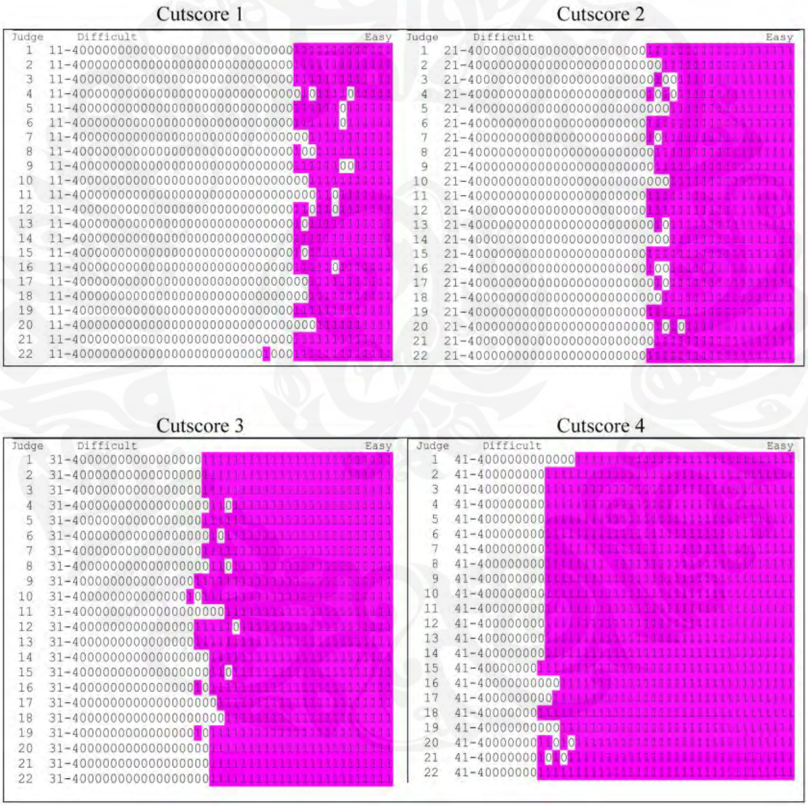
Figure 5: Cutscores/Criterion Levels as Applied to Examinee and Item Distributions (Grammar Subtest)



```
      EXAMINEES MAP OF
   5              .



                  -



                 .#

   3
                 .#
                  .  |
                .##

                .##
                  .
   2          .####

              .####
           .#######
             .#####
                     |         5GL -L4
            .######
          .########  |
   1     .#########
          .########  |  31GL4
        .##########  |  13GL1
        .########## M|     -L
       .###########  |  17GL2     22GL1
      .############  |
      .############  |     -L
         ##########  |  12GL4       -L4
   0    .##########  +M 19GL2
        .##########  |  14GL2
           .####### S|  16GL4     2GL2       36GL1     2
           ########  |  10GL4
             .#####  |  15GL4
             .#####
                ###  |
                 .#
  -1             .#  +S 11GL1     1GL1
                  . T|

                     |     -L
                  -
                     |
                     |

  -2                 +     -L

 EACH '#' IS
```

Figure 6 gives a simulated judging dataset involving the same grammar test items. The items for each of the cutscore are selected based on their empirical calibrations. About 25% of the items are selected to represent the first cutscore. Less than 50% of the items are selected to represent the second cutscore. Cutscores 3 and 4, on the other hand, are represented by approximately 60% and 80% of the items respectively.

Figure 6: Simulation of Ratings of Essential Items by Cutscore (level)



Cutscore 1

```
Judge      Difficult                            Easy
  1   11-400000000000000000000000000001111111111111
  2   11-400000000000000000000000000001111111111111
  3   11-400000000000000000000000000001111111111111
  4   11-400000000000000000000000000000101111011111
  5   11-400000000000000000000000000001111110111111
  6   11-400000000000000000000000000001111110111111
  7   11-400000000000000000000000000000011111111111
  8   11-400000000000000000000000000001001111111111
  9   11-400000000000000000000000000001111110011111
 10   11-400000000000000000000000000000011111111111
 11   11-400000000000000000000000000000001101111111
 12   11-400000000000000000000000000001101101111111
 13   11-400000000000000000000000000001011111111111
 14   11-400000000000000000000000000001111111111111
 15   11-400000000000000000000000000001011111111111
 16   11-400000000000000000000000000001111101111111
 17   11-400000000000000000000000000000011111111111
 18   11-400000000000000000000000000000011111111111
 19   11-400000000000000000000000000001111111111111
 20   11-400000000000000000000000000000001111111111
 21   11-400000000000000000000000000001111111111111
 22   11-400000000000000000000000010001111111111111
```

Cutscore 2

```
Judge      Difficult                            Easy
  1   21-400000000000000000000001111111111111111111
  2   21-400000000000000000000000011111111111111111
  3   21-400000000000000000000000100111111111111111
  4   21-400000000000000000000001010111111111111111
  5   21-400000000000000000000000001111111111111111
  6   21-400000000000000000000001111111111111111111
  7   21-400000000000000000000001011111111111111111
  8   21-400000000000000000000000111111111111111111
  9   21-400000000000000000000000111111111111111111
 10   21-400000000000000000000000001111111111111111
 11   21-400000000000000000000001111111111111111111
 12   21-400000000000000000000001111111111111111111
 13   21-400000000000000000000000101111111111111111
 14   21-400000000000000000000000001111111111111111
 15   21-400000000000000000000001111111111111111111
 16   21-400000000000000000000001001111111111111111
 17   21-400000000000000000000000101111111111111111
 18   21-400000000000000000000000011111111111111111
 19   21-400000000000000000000001111111111111111111
 20   21-400000000000000000000000101011111111111111
 21   21-400000000000000000000000111111111111111111
 22   21-400000000000000000000001111111111111111111
```

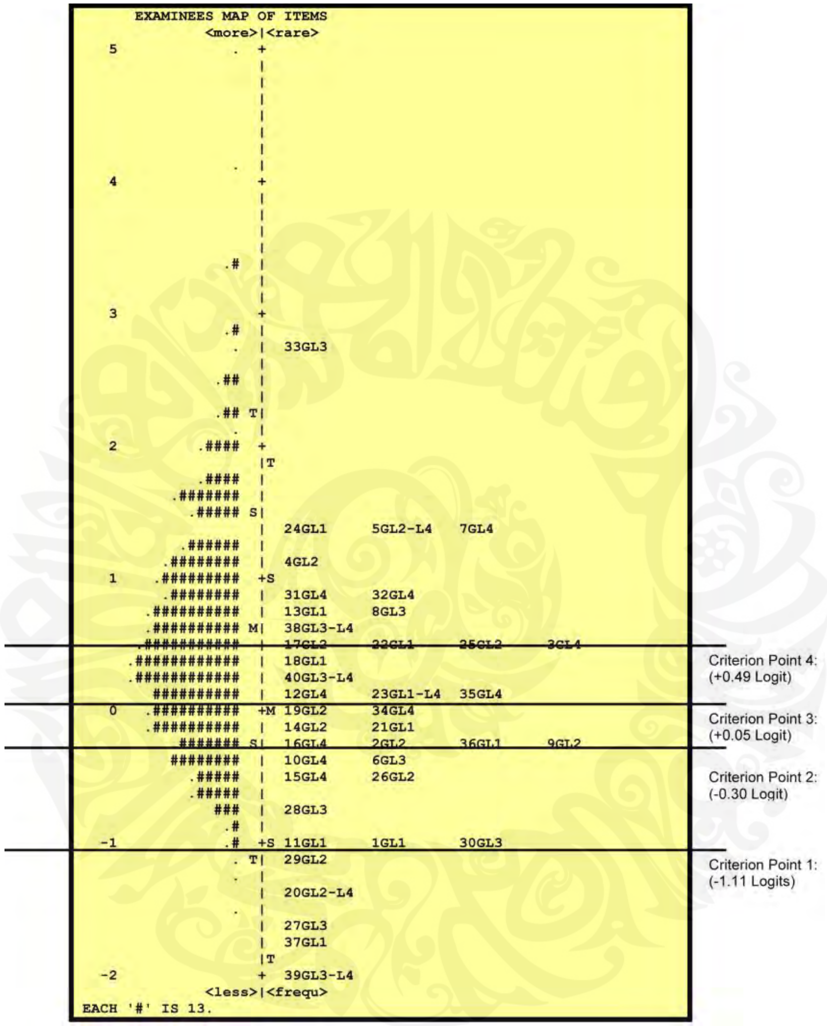
Cutscore 3

```
Judge      Difficult                            Easy
  1   31-400000000000000001111111111111111111111111
  2   31-400000000000000001111111111111111111111111
  3   31-400000000000000001111111111111111111111111
  4   31-400000000000000000110111111111111111111111
  5   31-400000000000000001111111111111111111111111
  6   31-400000000000000000101111111111111111111111
  7   31-400000000000000001111111111111111111111111
  8   31-400000000000000000110111111111111111111111
  9   31-400000000000000011111111111111111111111111
 10   31-400000000000000101111111111111111111111111
 11   31-400000000000000000001111111111111111111111
 12   31-400000000000000011111011111111111111111111
 13   31-400000000000000011111111111111111111111111
 14   31-400000000000000000111111111111111111111111
 15   31-400000000000000000110111111111111111111111
 16   31-400000000000000010111111111111111111111111
 17   31-400000000000000000011111111111111111111111
 18   31-400000000000000000001111111111111111111111
 19   31-400000000000000010111111111111111111111111
 20   31-400000000000000000111111111111111111111111
 21   31-400000000000000000111111111111111111111111
 22   31-400000000000000000111111111111111111111111
```

Cutscore 4

```
Judge      Difficult                            Easy
  1   41-400000000000011111111111111111111111111111
  2   41-400000000111111111111111111111111111111111
  3   41-400000000111111111111111111111111111111111
  4   41-400000000111111111111111111111111111111111
  5   41-400000000111111111111111111111111111111111
  6   41-400000000111111111111111111111111111111111
  7   41-400000000111111111111111111111111111111111
  8   41-400000000111111111111111111111111111111111
  9   41-400000000111111111111111111111111111111111
 10   41-400000000111111111111111111111111111111111
 11   41-400000000111111111111111111111111111111111
 12   41-400000000111111111111111111111111111111111
 13   41-400000000111111111111111111111111111111111
 14   41-400000000111111111111111111111111111111111
 15   41-400000001111111111111111111111111111111111
 16   41-400000000001111111111111111111111111111111
 17   41-400000000011111111111111111111111111111111
 18   41-400000001111111111111111111111111111111111
 19   41-400000000001111111111111111111111111111111
 20   41-400000001101011111111111111111111111111111
 21   41-400000001010111111111111111111111111111111
 22   41-400000001111111111111111111111111111111111
```

Figure 7: Cutscores/Criterion Levels Based on the Simulation Data



```
          EXAMINEES MAP OF ITEMS
                  <more>|<rare>
    5                 .  +
                         |
                         |
                         |
                         |
                         |
                         |
                      .  |
    4                    +
                         |
                         |
                         |
                         |
                     .#  |
                         |
                         |
    3                    +
                     .#  |
                      .  |  33GL3
                         |
                    .##  |
                         |
                    .## T|
                      .  |
    2             .####  +
                         |T
                  .####  |
               .#######  |
                 .##### S|
                         |  24GL1      5GL2-L4    7GL4
                .######  |
              .########  |  4GL2
    1        .#########  +S
              .########  |  31GL4      32GL4
           .##########   |  13GL1      8GL3
           .########## M|  38GL3-L4
         .###########    |  17GL2      22GL1      25GL2      3GL4
         .############   |  18GL1
         .############   |  40GL3-L4
            ##########   |  12GL4      23GL1-L4   35GL4
    0      .##########   +M 19GL2      34GL4
           .##########   |  14GL2      21GL1
              ####### S|  16GL4      2GL2       36GL1      9GL2
              ########   |  10GL4      6GL3
                 .#####  |  15GL4      26GL2
                 .#####  |
                    ###  |  28GL3
                     .#  |
   -1                .#  +S 11GL1      1GL1       30GL3
                      . T|  29GL2
                      .  |
                         |  20GL2-L4
                      .  |
                         |  27GL3
                         |  37GL1
                         |T
   -2                    +  39GL3-L4
                  <less>|<frequ>
 EACH '#' IS 13.
```

The resulting criterion levels (cutscores) from the simulation data are presented in Figure 7. On the whole, the criterion levels (cutscores) are more reasonable. Nonetheless, the first criterion level (cutscore) is still rather low. This perhaps could be due to the off-target items at the bottom of the measured scale and the small selection of essential items. The second and third criterion levels (cutscores) are closer together. The fourth cutscore is a little underestimated.

Discussion and Conclusion

This Many-facets Rasch analysis in combination with the OSS procedure has some clear advantages. First, it facilitates efficiency in identification of judges' internal inconsistency. In using this procedure, inconsistent judges can be identified in the same process in which the criterion levels (cutscores) are calibrated. If it is decided that the ratings of inconsistent judges are to be excluded from the computation of the final criterion levels (cutscores), subsequent computation of the criterion levels (cutscores) can be easily and quickly processed.

The second advantage of utilizing this approach pertains to judgment of essential items. Items that are misjudged as difficult or easy can be easily identified through the use of fit statistics and displacement values. Third, as the Many-facets Rasch analysis computes the error of measurement with regard to judges' ratings for each criterion level (cutscore), this can be used in the adjustment of the final criterion levels (cutscores), if so desired.

In relation to situations where multiple criterion levels (cutscores) are necessary, statistical significance of criterion level (cutscore) separation can also be clearly established. This approach to computing judges' ratings also allows for the investigation of judge-item interaction. Through bias analysis, unexpectedly harsh or lenient ratings with regard to any particular items by particular judges can be identified. These "idiosyncratic ratings can be intercepted and, if necessary treated as "missing" without disturbing the validity of the remainder of the analysis".[38] Alternatively, feedback can be given to the judges in question for improvements in the judging process.[39]

The findings of this study also underscore the importance of a clear understanding of the measured construct. As criterion levels

[38] Linacre, *Many-Facet Rasch Measurement* . . .

[39] Ibid.

(cutscores) and performance standards are set on test/s, the quality of the test/s used is bound to impact the derived criterion levels (cutscores) to some extent. Therefore, it is critical that the quality of the test/s used in a standard setting study/ process is carefully examined. This is particularly important when model-based methods are involved as explained by Kane:[40]

> The model-based, theoretical interpretation is considerably richer than a simple generalization from performance on a sample of tasks to expected performance on the universe of tasks from which the sample is drawn, and as a result, requires more evidence for its support. In particular, the validity argument for model-based, theoretical interpretations will require evidence for the validity of the theory of performance as well as evidence that the assessments can be interpreted in terms of the theory. That is, an interpretation of test scores in terms of a theoretical model depends on evidence for the model and for the relationship between the observed scores and terms in the model.

Judge competency must also be given due attention. Based on a study involving judge competency[41], it is suggested that judges should not only be trained to perform the judgment task but also should be trained in "the domain content for which they are to set the competency standard".

Fairness and equitable due process are fundamental for accountability. It is aptly asserted that in situations where there are sanctions for falling short, the use of performance standards needs to be handled with great care and justice[42]. Our measurements are not free of error; hence, the most appropriate method should be chosen to minimize the negative impact of any decisions made. Although the

[40] Michael Kane, "Validating High-Stakes Testing Programs," *Educational Measurement: Issues and Practice* 21, no. 1 (2005): 32.

[41] Lei Chang et al., "Does a Standard reflect Minimal Competency of Examinees or Judge Competency?" *Applied Measurement in Education* 9, no. 2 (1996): 161–173.

[42] Linn and National Center for Research on Evaluation, Standards and Student Testing, "Performance Standards: Utility for . . ."

lack of objectivity due to the judgmental nature of the standard setting process renders cutscores or performance standards somewhat arbitrary, they need to be made. Therefore, the setting of standards has to be handled with great prudence and a consciousness of what it entails and the stakes involved, as appropriately argued by Popham,[43]

> …when human beings apply their judgmental powers to the solution of problems, mistakes will be made. However, the fact that judgmental errors are possible should not send educators scurrying from such tasks as the setting of standards. Judges and juries are capable of error, yet they do the best job they can. Similarly, educators are now faced with the necessity of establishing performance standards, and they, too, must do the best job they can. That educational performance standards need to be set in order for instructional decisions to be made is indisputable. That those standards will, in the final analysis, be set judgmentally is equally indisputable. However, that all judgmental standards must be arbitrary is decidedly disputable. Approaching the standard-setting task seriously, taking full cognizance of available data and the preferences of concerned constituencies, need not result in an enterprise that is arbitrary and capricious. On the contrary, the act of standard-setting can reflect the very finest form of judgmental decision-making.

Islam has taught people to be just in all aspects of their lives, and make their best efforts to come to the right decision based on valid measures and observations. Justice is seen as one of the moral virtues and attributes of the human personality that must be upheld[44]. Abdullah and Junaid add that justice "represents moral rectitude and fairness, since it means things should be where they belong."[45] Therefore, people must observe justice and fairness in their measures and transactions in their daily life. This could be applicable to

[43] Popham, *Modern Educational Measurement:* . . ., 372.

[44] Mohammad Abdullah and Mohammad Nadvi, "Understanding the Principles . . ."

[45] Ibid. 275.

educational settings where accurate decisions have to be made to grant each person his/her rightful dues (i.e. to put the right person in the right place and to give each one his/her rights). This could not be achieved without using reliable and valid instruments/ scales. In other words, the scales or instruments used to measure for example, students' performance on a test, should ensure and provide accurate/reliable interpretations on the measured construct or trait.

AL-SHAJARAH

Special Issue

Contents

WoS-Indexed under *Arts & Humanities Citation Index, Current Contents/Arts and Humanities* and **Scopus**

ISSN 1394-6870

9 771394 687009