James Madison University
JMU Scholarly Commons

Department of Graduate Psychology - Faculty Scholarship | Department of Graduate Psychology

6-2002

Equating Multiple Forms of a Competency Test: An Item Response Theory Approach

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Follow this and additional works at: http://commons.lib.jmu.edu/gradpsych

 Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
DeMars, C. (2002, June). Equating multiple forms of a competency test: An item response theory approach. Paper presented at the annual meeting of the Association for Institutional Research, Toronto.

This Presented Paper is brought to you for free and open access by the Department of Graduate Psychology at JMU Scholarly Commons. It has been accepted for inclusion in Department of Graduate Psychology - Faculty Scholarship by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.

Equating Multiple Forms of a Competency Test: An Item Response Theory Approach

Christine DeMars

James Madison University

(2002, June). Paper presented at the annual meeting of the Association for Institutional Research, Toronto.

Abstract

A competency test was developed to assess students' skills in using electronic library resources. Because all students were required to pass the test, and had multiple opportunities to do so, multiple test forms were desired. Standards had been set on the original form, and minor differences in form difficulty needed to be taken into account. Students were randomly administered one of six new test forms; each form contained the original items and 12 pilot items which were different on each form. The pilot items were then calibrated to the metric of the original items and incorporated in two additional operational forms.

Equating Multiple Forms of a Competency Test: An Item Response Theory Approach

The items on a test are intended to be a sample of a content domain. This allows educators to make inferences about what students know in the broader domain based on how they did on a particular sample of test items. If students learn the particular content of items, generalizations from performance on those items to the domain will no longer be valid. This can happen when the same test, or part of the same test, is administered to multiple sections of a course (not uncommon in general education and introductory courses) and students in earlier sections tell others about the test questions. It can also happen when the same test is used for multiple re-take opportunities.

To combat this issue, large standardized testing programs have multiple test forms. For example, the SAT and ACT are different at each administration. The GRE is offered continuously, rather than at discrete administration times, but every student takes different combinations of items. Tests aligned with a particular curriculum, such as the Advanced Placement tests, can also have multiple test forms.

Test items vary in difficulty, and if items are randomly assigned to test forms the forms will differ somewhat. This can be adjusted for by equating different forms to a common scale. This is one reason most standardized tests report some sort of standard score (such as the GRE scale from 200 to 800) instead of a percent-correct score; a 76% on one form may be equivalent to a 74% on another more difficult form, so it is equitable to report the same standard score for both scores.

There are two basic classifications of equating procedures: classical and item response theory (IRT) (Crocker & Algina, 1986, chapter 20; Kolen & Brennan, 1995, chapter 2; Petersen, Kolen, & Hoover, 1989). Classical procedures operate at the total-test score level. The most common classical procedures are linear equating and equipercentile equating. In their simplest

form, these procedures are used when either the same group of students or equivalent groups of students take two forms of a test (more complex designs use an anchor test, with items that are common to both forms, to account for differences between the groups when non-equivalent groups are used). In linear equating, a linear transformations (multiplication and addition of constants) is used such that after the transformation scores on both forms have the same mean and standard deviation. In equipercentile equating, the cumulative distributions of the two forms are plotted, and the score from one form is transformed to the score on the other form which has the same percentile rank. In contrast, item response theory procedures operate at the item level rather than the total test level. In a process referred to as calibration, the characteristics of the items (at a minimum, their difficulties) are estimated such that at any point on the ability continuum the probability of correct response can be predicted. This allows for quick equatings of different combinations of items, unlike classical equating where the entire equating process needs to be calculated whenever items are added or subtracted to a form. It also allows for each examinee to have a unique form.

Item response theory works by using a non-linear, logistic function to describe the probability of correct response (or the probability of responding at each score level if the item allows for partial credit). The function includes parameters for item difficulty, and sometimes for item discrimination (the probability increases steeply for discriminating items), and guessing (even very low-ability examinees may have some non-zero probability). These item parameters can be estimated through specialized software; BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used in this study. After the item parameters are estimated, standardized scores can be estimated for the examinees. The scaling of the item parameters is arbitrary; typically, the item parameters are scaled such that the examinees' scores have a mean of zero and standard deviation of one (the scores can then easily be transformed to some other scale for reporting,

such as a mean of 500 and standard deviation of 100 for the GRE). If the groups used to calibrate the items are not equivalent, the item parameters from one form will need to be adjusted to put them on the same scale as the item parameters from the other form, but once that adjustment is made any combination of items can be assembled into a new form.

In this study, multiple test forms were equated for a test designed to measure college students' skills in using electronic library resources. At one public university, students are required to pass this test before beginning the sophomore year, because the faculty view these skills as necessary for advanced coursework. Students take the test at a computer lab and may attempt it whenever they feel ready, and they may re-take the test as often as they want (with the restriction that they may take it only once each day). Extensive on-line tutorials (intended to be incorporated in general education courses) as well as group workshops led by reference librarians are available to help students learn the material. Because students may take the test multiple times if they wish, and because students take the test at different times, there are opportunities for students to learn what particular items are on the test and it is reasonable to think some students would learn what was needed to answer these items without necessarily learning the broader content domain. During the 1999-2000 and 2000-2001 academic years, there was only one test form. During the fall of 2001, new test items were piloted, and this study will describe the process of equating these new items to the original test.

Method

Instrument

The assessment instrument, titled the Information Seeking Skills Test (ISST) was written by reference librarians at the university. To teach students library skills, the reference librarians created a series of on-line lessons. Each module begins by listing the lesson's objectives and ends with a short self-quiz. Instructors in foundational general education courses incorporate

these lessons as homework assignments. The ISST is based on the objectives covered in the lessons and is similar to the self-quizzes. The base form (used during 1999-2000 and 2000-2001) had 53 items. The test scores of first-time test takers had a coefficient alpha reliability of .70 in the 2001-2002 academic year (reliability was higher, .87, when it was first piloted in a context where students were tested at a fixed time rather than when they felt prepared, because the standard deviation of scores was much larger in that context).

Procedure

In 2001-2002, 72 new items were pilot-tested. There were six test forms, and the first time a student took the test one of the six forms was randomly selected (if the student took the test again at a later time, a different form was selected). Each of the forms contained the 53 items from the base form as well as 12 pilot items. Because students receive their scores at the time of testing (scoring could not wait until enough data was collected for equating), scores were based on the 53 original items. The items were administered in random order and students did not know which items were the pilot items. Data were available from about 500 first-time examinees per form (sample size ranged from 448-524).

A passing score had previously been set on the metric determined by the base year group (353 sophomores enrolled in related general education courses during the 98-99 academic year). Thus, all the new items were calibrated such that the item parameters would be on this same metric. The pilot items and the operational items were calibrated concurrently using BILOG-MG. Data from the base group were included, and the estimated population distribution for this group was defined to have a mean of zero and a standard deviation of one (equivalent to

procedures used in the originally scaling)[1]. A 3-parameter logistic model was used for the calibration. The model used was:

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta_j - b_i)}}{1 + e^{1.7a_i(\theta_j - b_i)}}, \quad (1)$$

where

$P_i(\theta j)$ was the probability that examinee j answers item i correctly, given examinee j's ability, θ,

b_i was the difficulty for item ,

a was the discrimination (slope) for item i, and

c_i was the lower asymptote.

An example of such a probability function (often called an item characteristic curve) is shown in Figure 1.

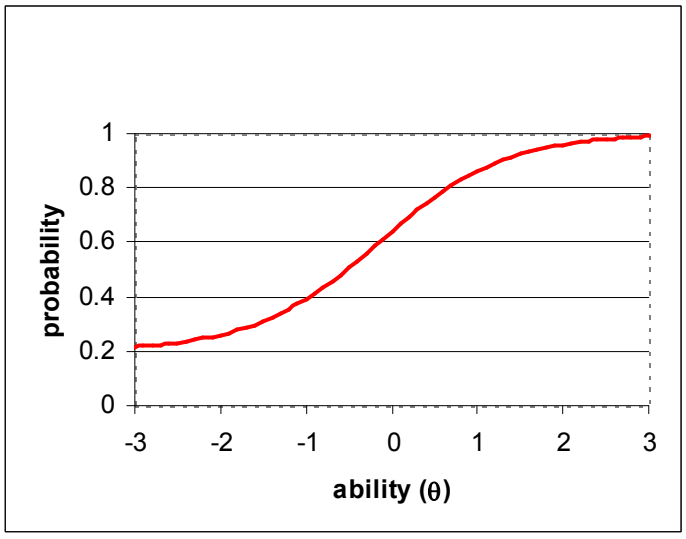

Figure 1: An item characteristic curve

Results and Conclusions

[1] the 2001-2002 students had a higher mean and a smaller standard deviation. Because students could take the test when they felt ready, they tended to score higher and in a smaller range.

Table 1 displays the expected number correct score on each set of 12 pilot items for several standard scores across the ability range. The standard scores are a linear transformation of the IRT scores, with a mean of 500 and standard deviation of 100 in the base year. The expected number correct scores (also called the IRT true score) are estimated by substituting in the IRT score in equation 1, estimating the probability of correct response, and summing these probabilities.

Table 1

Expected Number Correct Scores on Pilot Items

Standard Score	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6
200	3.1	3.0	3.0	2.5	2.8	3.0
250	3.5	3.3	3.4	2.6	3.0	3.4
300	4.0	3.9	3.9	2.8	3.3	4.0
350	4.5	4.8	4.6	3.3	3.9	4.9
400	5.1	5.9	5.5	4.0	4.7	5.9
450	5.8	7.2	6.6	5.2	5.9	6.8
500	6.9	8.5	7.9	6.6	7.3	7.7
550	8.2	9.8	9.3	7.9	8.8	8.7
600	9.3	10.6	10.2	8.8	10.0	9.4
650	10.3	11.1	10.9	9.6	10.8	9.8
700	11.0	11.4	11.3	10.2	11.3	10.2
750	11.5	11.6	11.6	10.7	11.7	10.5
800	11.8	11.8	11.8	11.1	11.8	10.8

From Table 1, we can see that the number correct scores do not always map to the same standard score. For example, an examinee with a score of 500 would be expected to score 6.6 on form 4 but 8.5 on form2 (of course, a full-length test would have more than 12 items so there would likely be smaller chance differences). If students were provided only with their standard scores, this would not present a problem. However, at the present time students are instead given number correct scores. Obviously, it would create confusion and resentment if two students had the same number correct score and one student passed and the other student failed

(understanding equating on a cognitive level and accepting it on an emotional level are different things when negative consequences are involved). Therefore, an attempt was made to shuffle items in the creation of new forms such that at the cut-score (an IRT of 0.13, or standard score of 513) the number correct score would be 42 on any form (the same as it was on the original form).

The plan was to have each of two new forms consist of 20 anchor items (also called common items--items that are the same on different forms) from the original items and 33 unique items chosen from among the pilot items (this would allow for discarding up to six of the pilot items if they could not be calibrated or had very low discriminations). This plan was modified somewhat given the availability of items in the different content and process categories. The test blueprint for the original form is shown in Table 2. The number of pilot items available in each category, after discarding five items which could not be calibrated (all examinees got one of the items right, and for the other four items BILOG-MG never reached stable parameters after 100 iterations, perhaps because very few students answered the items correctly), is shown in Table 3.

In order to be able to use as many of the pilot items as possible in the new forms, the test blueprint was changed, such that the numbers of items in the six categories (basic reference, database, internet, ethics, knowledge, and application) were the same as in the original blueprint, but the numbers in the cells created by crossing process with content changed slightly. Also, two items were cross-classified in two content areas, so two additional items were added to database (without changing the total number of items). Some of the pilot items, rather than just the original items, were also used as anchor items. The number of anchor items in each cell was based mainly on the availability of pilot items; where there were enough unique pilot items to fill a cell, no or few anchor items were chosen for that cell. When there were enough pilot items in a given cell for one, but not both, forms, the same pilot items were used on both forms (making them anchor items). While it would have been better from a measurement standpoint to select

items in each cell for the anchor so that the anchor was as representative of the complete test as possible, from a practical standpoint it was more important to use more new items and replace some of the items that had been used for several years. The resulting blueprint is shown in Table 4.

Table 2

Original Test Blueprint

	Process	
Content	Knowledge	Application
Basic Reference	18	1
Database Searching	12	8
Internet	4	7
Ethics	3	0

Table 3

Categorization of Pilot Items

	Process	
Content	Knowledge	Application
Basic Reference	20	0
Database Searching	25	14
Internet	7	1
Ethics	3	0

Table 4

New Test Blueprint

	Process	
Content	Knowledge	Application
Basic Reference	6 old anchor, 10 new unique, 2 old unique	1 old anchor
Database Searching	11 new unique	4 old anchor, 7 new unique
Internet	2 old anchor, 6 new anchor	1 old anchor, 1 new anchor, 2 old unique
Ethics	3 new anchor (2 duplicate Internet)	0

note: 'old' refers to original items, 'new' refers to pilot items

For the basic reference knowledge, database knowledge, and database application cells, an algorithm was written to find all possible ways of splitting the available pilot items into two forms (to simplify this, the three database knowledge items with the lowest discriminations were left out of the pool, so that there were exactly the number of pilot items needed for two forms). The split which minimized the difference (between the two forms) in expected number correct score and information at the cut-point was chosen (information in IRT is analogous to reliability, except that the information varies at different scores). In the basic reference application, internet knowledge, or ethics cells all items (original and pilot) were anchor items; when there were multiple items to choose from, ultimately the easier items were chosen (at first, the more discriminating items were chosen, but then revisions were made because the new forms were more difficult than the old forms). Similarly, the old unique items in the basic reference knowledge cell and the internet application cell were selected from the least difficult items when it became apparent that the new test forms were more difficult than the original form.

The resulting test forms were very close to each other in difficulty and information. At the cut-point, the expected number correct score was 39.5 on form A and 39.6 on form B, and the information was 16.8 and 16.7 respectively (corresponding to standard errors of measurement of .24), . Unfortunately, the aim had been to reach an expected number correct score of 42 on both forms. The only way to do this would be to increase the overlap of items on the two forms, or possibly to adjust the blueprint so that cells with easier items were more heavily represented. Neither of these were desirable, but two more test forms were created to meet the criteria of an expected number correct score of 42 at a theta of 0.13 (standard score of 513), by replacing 6 of the unique pilot items (3 on each form) and 1 pilot anchor item with 4 original anchor items in the same content areas. This meant the tests shared 25 anchor items instead of 21, and 7 fewer of the pilot items were utilized. This alternative will be presented to the content committee responsible for this test and they can choose which set of forms to use, based on which criteria are more important (minimizing form overlap, using as many new items as possible, or keeping the cut-score at the same number-correct score as the previous year).

While the main objective was for the expected number correct scores to be as equal as possible at the cut-point, it would be ideal if the expected number correct scores were equal at each standard score. Table 5 shows the expected number-correct score on each form across a broader range of standard scores. The number-correct scores were within half a score point (or nearly) at each of these standard scores, which is favorable.

Table 5

Expected Number Correct Scores on Each Form

Standard Score	Form A	Form B
200	12.9	13.5
250	14.8	15.4
300	17.8	18.4
350	22.1	22.5
400	27.5	27.6
450	32.9	32.9
500	38.2	38.2
550	42.9	43.2
600	46.4	46.7
650	48.9	49.2
700	50.6	50.9
750	51.7	51.9
800	52.3	52.5

When one is choosing a limited number of items from a larger item bank, more complex algorithms for test assembly have been proposed (see van der Linden, 1998, for some examples). The relatively simple procedure for this study was used because nearly all of the available items were to be assigned to one form or another, which changed the nature of the task from finding an optimal set of items to distributing all of the items into, as far as possible, equally difficult sets (at least at the cut-point).

In this study, the primary purpose was to assemble two new test forms so that students would not take all of the same items if they were re-tested and to reduce disclosure of items. Eventually, the goal for the ISST is to have a larger item pool, so that computer adaptive testing (CAT) can be utilized. With CAT, a preliminary IRT score will be estimated after the student has taken several items. Then, the next item will be selected so that the item difficulty matches the student's estimated proficiency score, the proficiency estimate will be updated and used to select the next item, and so on. Each student will have a number-correct score of about 50% with items chosen in this fashion, but the IRT scores will adjust for differences in item difficulty. This will

mean that number-correct scores will not be reported and will eliminate the need to build the test forms such that they map to the same number-correct score at the cut-point. However, additional items (at least 200) need to be written, pilot-tested, and calibrated before CAT can be implemented for the ISST.

Conclusion and Implications for Institutional Research

Program assessment is increasingly becoming a responsibility of institutional researchers (with some exceptions, such as this university, where assessment is the responsibility of a related unit). Conducting large-scale assessment of university-wide programs requires (and allows) the use of more sophisticated techniques than can be used in classroom assessment. Security issues are more of a concern when the assessment involves greater numbers of students; creating multiple forms of a test can help address these concerns. If students are held accountable for demonstrating a certain level of proficiency, and are re-tested until they can do so, validity issues become a concern because a student could memorize specific test questions without improving on the construct underlying the test. Multiple test forms can help address this concern also. But when more than one test form is used, the forms must be equated. IRT provides a way to equate the forms; once all the items are calibrated to the same metric, the IRT or standardized scores are equivalent. If number-correct scores are to be reported, as in this study, further work needs to be done by assigning items to forms in such a way that the number-correct scores are as equivalent as possible.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practice*. New York: Springer.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1993). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 221-262). Phoenix, AZ: Oryx Press.

van der Linden, W. J. (1998). Optimal test assembly [Special issue]. *Applied Psychological Measurement, 22*(3).

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). BILOG-MG [computer software]. Chicago: Scientific Software International.