The Advantages of Transitioning from a Set-Form Testing Approach to Test Item Bank Methodology

Mark Duane STAFFORD

English Education Center, Ehime University

Introduction

The English Education Center (EEC) at Ehime University, founded in 2001, began developing a common test program for the university's required general education English courses during the 2009-2010 academic year. The goals of the common test program were to establish minimum English proficiency levels for students to attain during their freshmen-year, to assess whether the goals and objectives of the program were being met, and whether students demonstrated improvement in their abilities over four semester-long skill-based English courses.

EEC courses were split into separate listening, speaking, reading, and writing courses in 2007. Listening and speaking are offered in the first semester and reading and writing in the second. Common tests for the speaking and writing courses are based on rubrics developed by members of the EEC, whose development underwent a number of iterations before being finalized into the current form which has remained in use for the past seven years.

The development of the listening and reading common tests, on the other hand, has been decidedly less expedient for a variety of reasons. First of all, substantial time and effort has been required in developing the current 200 listening test questions[1] and 120 reading test questions compared to the rubric-based tests. While scoring descriptors within the rubrics can be quickly and easily revised on a single form, each multiple-choice-based question must be piloted, examined for effectiveness (by employing classical test theory [CTT] item analysis statistics) (Brown, 1996), revised based on such data, and re-piloted, re-examined, and revised again in an ongoing process until desired results are obtained. In addition, each individual test question must be evaluated with regard to its contribution to the overall difficulty of the test — i.e. the mean score. Of the EEC's listening and reading A, B, C, and D sets, the A forms have gone through the developmental process at least six times and are therefore the most sophisticated in terms of statistical quality. The A sets still need to undergo a few more developmental cycles and the B, C, and D sets even more, requiring a number of years before all of the listening and reading tests reach the standards of quality that the EEC desires.

Once the battery of set-form tests are fully developed, however, there is no guarantee that they will continue to reflect the current state of affairs of incoming students at Ehime University. Indeed, the common tests are currently being formed to achieve an overall average score of 75% among 1,800 students from six faculties. Due to future social and educational policy changes and possible redistribution of Ehime University's student population among its six faculties, the immutable disposition of the set-form approach is a distinct disadvantage.

In addition, since the common listening and reading tests presently account for 30% of students' final grades they can be considered as high-stakes examinations where security is a fundamental concern. Whether accidental or incidental, any public release of a common, set-form test would render it useless to future application and the energies mustered in developing it would seem sorely wasted.

Endeavoring to overcome these developmental obstacles and to add enhanced features to the listening and reading common test program, members of the EEC secured an Ehime university Good Practice grant in the spring of 2015 to develop a test Item Bank (TIB). Making the transition from the set-form approach to the

1) The terms test "question" and test "item" are used interchangeably in this article.

TIB is a logical and necessary step in the evolution of the EEC's common test program. This paper describes the advantages of transitioning from a set-form, CTT approach to a TIB.

What Is A Test Item Bank ?

A test item bank is essentially a large pool of questions that can be selectively extracted to form a test unique to each administration. Once the questions are piloted and improved under CTT, the entire bank of questions is given to a large number of persons with diverse abilities. A Rasch analysis is then performed using specialized statistical software — Winsteps 3.81.0 (Linacre, 2015) in our case. Items that are found not to fit the measurement objective well may be revised, or rejected from inclusion in the TIB (Bond and Fox, 2007). Items assessed as having good fit are included in the TIB and catalogued according to their difficulty relative to the average person's ability as determined by the Rasch analysis.

Once the difficulty of each item is known, individual questions (or blocks of questions as with the common listening and reading tests) can be withdrawn from the bank to form a test with a predictable and very accurate overall average difficulty (mean score).

Item Response Theory (IRT) was selected to be employed in the design and analysis of the TIB because it is based on establishing a model that specifies the probability of observing each response option to an item as a function of the target trait being measured by the assessment, which is often a knowledge, skill, or ability. In testing situations where items are scored as correct or incorrect, IRT specifies the probability of a correct response to an item as a function of ability. (The University Of North Carolina At Greensboro, 2015)

In addition, the one-parameter model of IRT, also known as Rasch, was chosen over multiple-parameter models because it ranks item difficulty the same for all respondents independent of ability and person ability independent of difficulty (Bond and Fox, 2007). Therefore, it becomes relatively easy to form an objective ranking of item difficulty with the information that Rasch analysis provides and to use such information for the TIB.

The Benefits Of A Test Item Bank

Although considerable effort is involved in writing questions and conducting the Rasch analysis for the TIB, the benefits of such labors are substantial. As mentioned earlier, the problems with the set-form approach can be solved and additional advantages can be added to the listening and reading common test programs through developing the TIB.

Time Savings

Significant time can be spared by creating the TIB in comparison to going through the tedious developmental cycles of the set-form approach. Once a significant number of questions is generated, improved through CTT, and a Rasch analysis is performed, misfits can be excluded from the bank. Only items which fit the measurement objective remain in the bank and their difficulty is known and recorded. Although the abilities of persons will change over time, the difficulty of the items will not. Therefore, the TIB development process only needs to be conducted once compared to the many years required for the set-form approach.

Security

In addition to the time savings that developing a TIB can offer, there is also a substantial increase in security. As stated above, the public release of a set-form common test renders it invalid for future deployment. In addition, the information on set-form tests can be transferred to future generations of students by word of mouth. The TIB solves both of these problems by allowing testing program supervisors to withdraw different questions from the bank for each administration. This essentially means that, given a relatively high number of items in the bank, no question will be used more than once during a certain number of years. Each test will be unique, yet also equivalent in difficulty to prior administrations, by using the information provided by Rasch analysis.

Adaptability

The advantages in adaptability of a TIB in relation to a battery of set-form tests are manifold.

First is the ability to use TIB questions in both criterion-referenced and norm-referenced contexts. The listening and reading common A, B, C, and D tests were designed exclusively for a criterion-referenced situation intended to measure whether students attained a certain level of listening or reading proficiency. In doing so, most of the set-form questions fall into a narrow range of

difficulty determined by the level of proficiency set by the EEC's curriculum and its testing committee. In which case, a 75% average score was desired to ensure that a majority of students pass the test (and subsequently the course). The 75% standard holds up very well with the entire population of the freshman class of nearly 1,800 (minus dropouts and students who exempted from the course).

However, when each faculty is taken into consideration, a different picture emerges. Due to contrasting admission standards regarding English proficiencies, each faculty demonstrates a mean score which is different from the others, varying from the medical faculty's 84.7% to the agriculture faculty's 71.92% (FB, classes of mixed faculties repeating the course is even lower at 60.73%) as can be viewed in Table 1 containing data from the common reading test administered during the 2014-2015 academic year. While individual faculty mean scores are lower or higher than the target 75%, the weighting of student populations among the faculties determaines the overall average.

Table 1

Faculty	Mean Score	N	Proportion	Weighted Mean
Agriculture	0.719	172	10%	0.070
Education	0.736	211	12%	0.088
Engineering	0.733	449	25%	0.186
FB (repeaters)	0.607	100	6%	0.034
Law & Lit	0.784	479	27%	0.212
Medical	0.847	109	6%	0.052
Science	0.767	251	14%	0.109
Total	-	1,771	100%	0.751

Although creating items within a narrow difficulty range was advantageous and valid within a criterion-referenced context, fixing most questions at a specific, relatively low difficulty level does not provide us with information about higher-ability students, not to mention students in the lower percentiles. In other words, most of the current test questions are too easy for higher level students and too difficult for lower level ones. Therefore, developing questions of a variety difficulties for the TIB will add needed latitude to the listening and reading common test program, while still allowing for a highly predictable desired overall mean score of 75%. Including some questions on both sides of the difficulty spectrum will provide more information about "non-average" students for curriculum development purposes and allow the TIB to be used for norm-referenced purposes such as placement within a level-streaming curriculum, admission into advanced courses, or in lieu of a standardized test such as the GTEC (Benesse Corporation, 2004).

Second, a TIB will allow the common test program to adapt to changing English proficiency levels of future incoming students at Ehime University. Whereas the set-form approach can be slowly adapted to increasing or decreasing levels, the process is somewhat akin to a dog chasing its own tail. The process of adapting the set-form format is quite slow which may lead to never being able to catch up with changing levels. The TIB, on the other hand, allows for a quick response the variation in levels by merely selecting items for the next administration that constitute an easier or more difficult test than the previous administration.

Fairness

Furthermore, greater fairness for students taking the listening and reading common tests can be achieved by transitioning from the set-form approach to a TIB. Even though two set-form tests may be developed to produce identical mean scores under CTT (an accomplishment which remains incomplete within the EEC), test administrators assume that each question is of equal difficulty in relation to all other questions on both set-forms for all test-takers, which is highly implausible. Although results of two "identical" tests may share the same mean, how that mean is achieved may be very different (See Table 2).

Table 2

Ability	Form X	Form Y
Higher	.950	.850
Mid	.700	.800
Lower	.600	.600
Average	.750	.750

In this highly simplified example, the means of Forms X and Y are the same, yet Form X is obviously skewed in favor of higher ability students. Were the two forms administered to different groups in consecutive years (not to mention at the same time) means observed for each test would likely be very similar, but the two forms could not logically be considered "identical", which raises issues of the fairness of employing set-forms under CTT practices.

TIB offers solutions to this dilemma. Since Rasch analysis assigns a difficulty rating for items in relation to person's abilities, we can accurately predict how difficult the collective items on a test will be for each ability level

using what's called Test Information Functions (TIF) (Bond and Fox, 2007, Yu, 2010). Moreover, Rasch analysis software conveniently allows the comparison, and hence, balancing of difficulty for two or more alternate forms using TIFs. Clearly, using a TIB and Rasch analysis would add greater fairness to the listening and reading common tests whether administered over a span of time or during a single session.

Accuracy and Predictability

Using Rasch modeling with a TIB also offers much greater accuracy and predictability over CTT practices. In a pilot study to be reported later, Rasch analysis was used to predict the target mean of .750 for the 2014-2015 academic year common reading test. Items were selected from both Reading A and B forms to constitute a collective mean of 0.745, according to Rasch modeling. The resulting mean of the actual test was 0.751, meaning that Rasch modeling was accurate in it's prediction of the actual mean to within 0.006 points. When the mean that would have been predicted under CTT, which was based on results from the alternate A and B forms administration for the 2012-2013 academic year, is compared to the actual mean, a difference of 0.027 points can be observed. Where CTT methodology would have predicted the mean with an error of nearly three percent, Rasch modeling predicted the actual mean to within about half a percentage point. Clearly, the TIB and Rasch modeling approach holds significant advantages over the set-form and CTT approach. (see Table 3).

Table 3

CTT Predicted Mean	0.778
Rasch Analysis Predicted Mean	0.745
Actual Mean	0.751

Drawbacks

While the advantages of a TIB are many, there are also a few drawbacks associated with employing the TIB. For one, Rasch analysis on all TIB items must be performed again if any major changes or additions to the system are made. In addition, unless multiple Rasch analyses are performed, items of poor initial quality must be thrown out of the TIB rather than revised and improved. This is perhaps not the most efficient way to utilize the substantial human resources required to build the TIB.

Conclusion

In addition to offering accelerated test program development, the TIB, Rasch methodology approach brings greater security, adaptability, fairness, accuracy and predictability compared to the set-form, CTT approach. Although creating a TIB initially requires substantial time and labor, the long-term benefits are clearly worth such efforts. In light of the substantial benefits that a TIB could bring compared to the set-form approach, the EEC would indeed be amiss if it were to pass on the opportunity to take a great evolutionary step in developing its common test program.

Writing of additional passages and questions for the listening and reading common test TIBs is currently under way. Piloting of the new reading questions is planned during the 2015-2016 academic fall semester, and listening questions during the early 2016-2017 academic spring semester. TIB questions will then be revised and improved under CTT procedures. Finally, TIB items will be added to the regular common test forms as pseudo questions which will not be calculated in students' final common test scores. Then the entire TIB will be analyzed using Rasch methodology, item difficulties will be determined, and the questions will be catalogued according to their type and difficulty estimates. If existing courses endure approaching curriculum reform efforts, official use of the TIB will begin during the 2017-2018 academic year. Even if the curriculum is drastically altered, the opportunity to deploy the TIB in one form or another will hopefully remain.

References

Benesse Corporation (2004). GTEC, Global Test of English Communication. Okayama, Japan : Benesse Corporation.

Bond, T. G. and Fox, C. M. (2007). Applying The Rasch Model : Fundamental Measurement In The Human Sciences. New Jersey : Lawrence Erlbaum Associates, Inc.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ : Prentice Hall.

Linacre, J. M. (2015). Winsteps® Rasch measurement computer program. Beaverton, Oregon : Winsteps. com.

Yu, Chong Ho. (2010). A simple guide to the item response theory (IRT) and Rasch modeling. Downloaded from http://www.creative-wisdom.com/computer/sas/IRT.pdf

The University Of North Carolina At Greensboro. (2015). Retrieved from http://erm.uncg.edu/oaers/methodology-resources/item-response-theory/