SRB

SHIKEN RESEARCH BULLETIN

Volume 17 • Number 2 • December 2013

Contents

JALT
Testing and Evaluation SIG

Foreword

Jeffrey Stewart
TEVAL SIG Publications Chair,
Shiken Research Bulletin *General Editor*

Season's greetings to all of you from the editorial team of *Shiken Research Bulletin*! Our holiday gift to you is another great issue of *SRB*, packed with useful articles and columns. In this issue's feature article, Jonathan Harrison and Ruth Vanbaelen apply J. D. Brown's approach to curriculum design to English communication courses. Rasch columnist Jim Sick compares and contrasts Classical Test Theory and Rasch Measurement Theory approaches to the important issue of rater agreement in assessment, shedding further light on the utility of the program Facets. Statistics columnist J. D. Brown provides practical (and crucial) advice for classroom teachers developing their own tests. Finally, software columnist Aaron Batty introduces `ltm`, a package for R that provides the functionality of professional IRT software that normally costs thousands of dollars—for free. Those of you who attended our IRT workshop at the Japan Language Testing Association Symposium this year will find it of particular interest, as it documents how to perform the model-fit comparisons that we covered as a group step-by-step.

I'm happy to end this year on a high note, because due to the demands of writing my PhD thesis, 2013 marks the end of my term as editor. I'm looking forward to reading Shiken as a TEVAL member in the new year.

Brown's approach to language curricula applied to English communication courses

Jonathan J. Harrison
jon.harrison@nihon-u.ac.jp
Nihon University, College of Science and Technology

Ruth Vanbaelen
ruth.vanbaelen@nihon-u.ac.jp
Nihon University, College of Science and Technology

Abstract

Brown's approach to designing and maintaining language curricula consists of six processes: needs analysis, objectives, testing, materials, teaching, and evaluation. This flexible approach was adapted for use in a program involving learners with a variety of needs and with various proficiencies. This systematic approach to curriculum design and how it was applied to English Communication courses between the 2009 and 2011 academic years at a private university in Japan are summarized. Results from the online testing program and the online surveys are given and how the course evolved is discussed. Results from 2010 indicate 3-24% achievement gains, student satisfaction at 98%, and student perceptions of learning on the rise. Brown's approach was useful for this two-course program.

Keywords: curriculum design, English, perception, systematic approach

Introduction

Language educators have an abundance of teaching methods from which to choose; well-known examples include grammar-translation, situational, functional, topical, skills, and tasks. Brown (1995) organized these into four categories of language teaching activities: approaches, syllabi, techniques, and exercises. Approaches, based on theories of language and language learning, are the ways of defining what and how the students need to learn, examples include the grammar-translation approach and the communicative approach. Syllabi are ways of organizing the course materials, including structural, topical, skills, and tasks. Techniques are the ways of presenting the materials and teaching. These include grammar demonstration dialogues, lectures on rules of language, and discussions. Exercises, for instance, fill-in, cloze, copying, and restatement, are the ways of practicing what has been presented. Teachers' approaches and theories may differ, and many teachers tend to use multiple approaches, different types of syllabi, and various techniques and exercises simultaneously based on the perceived needs of the learners in their classrooms in order to effectively and efficiently help them learn.

Brown's approach to designing and maintaining a language curriculum draws from various models and is a systems approach which allows for logical program development. Brown's view is that curriculum development is ongoing as it is "a series of activities that contribute to the growth of consensus among the staff, faculty, administration, and students" (p.19). The approach consists of six interconnected processes: needs analysis, objectives, testing, materials, teaching, and evaluation. Briefly stated needs analysis for a particular institution is the systematic collection and analysis of information regarding what is necessary to satisfy the students' language learning requirements. Objectives, meaning precise statements regarding the skills and content the students should master to reach larger goals, must be set. From the objectives criterion-referenced tests should be made to measure learning, and norm-referenced tests should be used to compare student performance. With the needs analysis, objectives, and tests in mind, materials can be adopted, developed, or adapted. Decisions regarding teaching should be made by the teachers, and it is best if the teachers are part of the curriculum design process and that they are

supported by the administration. Evaluation, meaning program evaluation, is an ongoing, systematic collection and analysis of all relevant information, gathered through all of the other processes, which is necessary to improve the curriculum and to assess its effectiveness.

Brown's approach is a framework within which the English Communication (EC) courses could be systematically evaluated and evolve based on learner needs. Although EC is not a curriculum but merely two similar course titles with eight subtitled course options, Brown's approach provides a framework with defined processes for information gathering, goal setting, evaluation of learning, materials development, teaching and learning, and overall program evaluation. This approach considers a curriculum to be a process which can change and adapt to new conditions. What follows is a discussion of how the authors adapted this curriculum model to EC courses.

This approach was chosen as it is flexible and allows for evolution and maturation through a systematic process. The authors were first year employees creating a program for implementation during their second year, and the learners involved had a range of needs and language proficiencies. The approach was applied to English Communication courses from 2009 to 2011 academic years at a science and technology university in Japan. In 2009 EC courses replaced English Conversation courses, broadening the focus to include communication in English through not only speaking but also through writing and nonverbal communications. These courses are required for students seeking teacher certification and elective courses for other second through fourth year students who must acquire ten foreign language credits for graduation, six of which must come from English courses. EC courses may account for up to two credits of the required ten credit curriculum. These new EC courses consist of four subtitled optional courses with two levels each, and the courses were implemented in two stages. Typically, multiple courses with related contents that support each other would be called a program; however, as students can only take a maximum of two of the eight subtitled courses, the term "course" will be used throughout this paper. The subtitles are as follows:

- Public Speaking 1 (PS1) provides students with practical experience to learn basic presentation skills and to write and give structured speeches.
- Public Speaking 2 (PS2) introduces students to scientific research presentation skills and focuses on learning to ask and respond well to questions.
- Traveling Abroad 1 (TA1) allows students to learn and practice basic survival English skills for communicating in different travel and home stay situations and to introduce aspects of Japan.
- Traveling Abroad 2 (TA2) gives students practice with language learning strategies and test-taking strategies for the TOEFL to prepare them for possible future study abroad.
- Business 1 (BU1) focuses on basic business communication skills.
- Business 2 (BU2) builds business communication, presentation and discussion skills.
- Cultural Appreciation 1 (CA1) focuses on learning to introduce aspects of Japan and researching basic information on other cultures.
- Cultural Appreciation 2 (CA2) explores cultural, social, and economical differences between Japan and various countries.

Implementation began in 2009 with two subtitled courses with two levels each, and in 2010, two more subtitled courses with two levels each began (Table 1): From the second semester of 2010, all of the subtitled EC courses were offered. These eight subtitled courses are divided into thirty-four EC course

time slots which are spread across two campuses. Each semester between 650 and 850 students registered for these EC courses (Office of Educational Affairs, 2009-2011).

Table 1. ***Schedule of Implementation of the Courses***

	First Semester	Second Semester
Until 2008	English Conversation I	English Conversation II
2009	English Communication I: *Public Speaking 1* *Traveling Abroad 1* *Traveling Abroad 2*	English Communication II: *Public Speaking 1* *Public Speaking 2* *Traveling Abroad 1* *Traveling Abroad 2*
2010	English Communication I: *Public Speaking 1* *Public Speaking 2* *Traveling Abroad 1* *Traveling Abroad 2* *Business 1* *Cultural Appreciation 1*	English Communication II: *Public Speaking 1* *Public Speaking 2* *Traveling Abroad 1* *Traveling Abroad 2* *Business 1* *Business 2* *Cultural Appreciation 1* *Cultural Appreciation 2*
2011	English Communication I: All	English Communication II: All

Methods of Design and Implementation

Methods of Needs Analysis

A needs analysis, as defined by Brown, is "the systematic collection and analysis of all subjective and objective information necessary to define and validate defensible curriculum purposes that satisfy the language learning requirements of students within the context." (p.36) He prescribes three systematic steps: Making basic decisions about the needs analysis, gathering information, and using the information. In short, the initial needs analysis information was gathered in 2008 through informal student surveys, conversations with students, and meetings with teachers and administrators. One merit of Brown's approach which the authors kept in mind was how the flexibility of the framework allows for continual data gathering and use. The EC course subtitles and basic descriptions had been set by administration, and the authors of this paper were given the task of establishing guidelines and administering these new courses with assistance from a third full-time EC teacher.

Methods of Objectives

Student language output expectations, general skill objectives for all courses and course-specific content objectives for each subtitled course, specifically topics and vocabulary were detailed by early 2009 by the three full-time native English-speaking teachers. In this paper, the term "Guidelines" will be used to refer to the output expectations and the objectives. The guidelines were provided to all EC teachers as a minimum of what should be taught in the courses, and they were updated annually.

Expected minimums of graded student output for writing were set at 300 words and 400 words for levels 1 and 2, respectively. Minimums for graded oral communications were 200 and 300 words for the respective levels. The 2011 goals and objectives for the eight subtitled courses were as follows.

Public Speaking 1 (Basic Presentations)

- Identify the parts of a speech: introduction, body, and conclusion. Many teachers use a basic five-paragraph presentation format with an introduction, three body paragraphs, and a concluding paragraph.
- Take notes on speeches and lectures to obtain the thesis, main points, and important details.
- Use notes to answer questions, write responses or opinions, ask questions, or have discussions.
- Write and present one or two structured speeches, including an introduction, body, and conclusion.
- Ask questions to presenters and answer questions regarding their own presentations.
- Presentation basics: proper posture, natural gestures, eye contact, and voice inflection (word stress, varied pace of speech, pausing, etc.)
- Understand and use styles of presentation introduction (a.k.a. attention grabbers, hooks), including a question, a statistic, an anecdote (story), a quote, humor/a joke, or a definition.
- Understand and use important vocabulary: introduction, body, conclusion, paragraph, sentence, attention grabber, thesis, topic sentence, primary and secondary support, transition, restate(ment), opinion, example.
- Write emails in a proper format, with a subject line, greeting + receiver's title and name, body (message), salutation, and sender's name in the proper order (e.g. for communications with the teacher about speech preparation, for peer review of other students' speeches, to thank a presenter, or to ask for more information).

Public Speaking 2 (Academic Presentations)

- Identify the parts of an academic speech: title, author, abstract, introduction, methods, results, conclusion, and references. Optional teaching points include: discussion, acknowledgements
- Choose, research, and present a topic.
- Quote and paraphrase sources and properly cite them.
- Take notes on speeches, the main points, and important details.
- Ask questions about others' research and respond to questions about their own research (e.g. expressing opinions, building an argument, etc.).
- Write emails in a proper format, with a subject line, greeting and receiver's title and name, body (message), salutation, and sender's name in the proper order (e.g. for communications with other students, researchers, presentation proposal submissions, etc.).
- Understand and use important vocabulary: research, quote, paraphrase, source, cite, title, author, presenter, introduction, abstract, methods, materials, results, discussion, conclusion, references

Traveling Abroad 1

- Communicate in English in different situations, specifically survival English for airports, hotels, restaurants, and shopping. Other situations are the decisions of the teacher.
- Learn to use vocabulary for speaking, listening, writing, and reading tasks in the following situations:

  Daily, basic conversations (possibly including greetings, introductions, classroom English, days of the week, months, numbers, and exchanging contact information.)

  Airports - important vocabulary: immigration, documents, visa, customs, gate, passport, declare, boarding, fasten/seat belt, embark, return, depart, tray/upright position, take off, land.

  Hotels - important vocabulary: reception(ist), reservation, check-in/out, luggage/baggage, smoking, non-smoking, wake up call, tip, room service, continental breakfast, complimentary, buffet, concierge, cost.

  Restaurants - important vocabulary: reservation, check, bill, credit (card), cash, pay, order, smoking, non-smoking, waiter, waitress, appetizer, anything else, bills, coins, change, total, receipt.

  Shopping- important vocabulary: price, cash, credit card, purchase, refund, exchange, discount, shoplifting, cashier, salesperson, fitting room, try on, take off, put on, window shopping, just looking, receipt, (currencies).
- Write emails in a proper format, with a subject line, greeting + receiver's title and name, body (message), salutation, and sender's name in the proper order.
- Focus on asking and answering basic questions: who, what, when, where, why, how, how much.

Traveling Abroad 2 (TOEFL)

- Practice skills for academic classes in English when studying abroad, specifically:
  a. listening to lectures
  b. taking notes
  c. expressing opinions
  d. participating in group discussions
  e. summarizing
  f. integrating ideas from multiple sources
- Write emails in a proper format, with a subject line, greeting and receiver's title and name, body (message), salutation, and sender's name in the proper order.
- Increase knowledge of and fluency with vocabulary. Focus on study skills and methods to build student vocabularies by 50 to 200 words used in academic settings.
- Learn at least two note taking skills and strategies, possibly including use of graphic organizers and outlining.
- Learn speaking, reading, writing, and listening skills for the TOEFL focusing on giving opinions, comparison/contrast, and cause and effect (as specified by ETS in a comparison of the old and new TOEFL tests at http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_at_a_Glance.pdf. Retrieved on Jan. 29, 2009.

Business 1

- Take and make phone calls. Students learn to make a simple call and take a simple message. This includes learning expressions, as well as structure of a company. Prepositions of time and place need to be reviewed.
- Write emails in a proper format, with a subject line, greeting and receiver's title and name, body (message), salutation, and sender's name in the proper order.
- Both form and answer questions about personality, studies, and the future in a job interview situation.
- Describe a product or a process. Students should learn vocabulary (verbs describing change) and transitions. Students could do simple descriptions/comparisons of products (e.g. cell phone functions, PC, cars, etc.)
- Take notes on speeches or mini-business meetings to obtain the gist and the most important information.

Business 2

- Take and make phone calls. Students learn to make a detailed call and take a detailed message. This includes learning expressions, as well as structure of a company. Prepositions of time and place need to be reviewed.
- Write emails in a proper format, with a subject line, greeting and receiver's title and name, body (message), salutation, and sender's name in the proper order. Classroom tasks could include: Suggesting a meeting and including an agenda, writing one's own opinion, replying to the received email.
- Both form and answer detailed questions about personality, studies, and the future in a job interview situation.
- Give a detailed description of a product or a process using graphs and data. Students should learn vocabulary (verbs describing change) and transitions. Students could do simple descriptions/comparisons of products (e.g. cell phone functions, PC, cars, etc.)
- Take notes on speeches or mini-business meetings to obtain important information and specific details. The study of inferred meanings could also be part of this course.

Cultural Appreciation 1

- Describe aspects of culture or cultural items.
- Give and ask for opinions with a focus on cultures and countries.
- Integrate information obtained from listening and reading into spoken or written product.

Cultural Appreciation 2

- Describe aspects of culture or cultural items.
- Give and ask for opinions with a focus on cultures and countries.
- Summarize listenings and readings.

- Integrate information obtained from listening and reading into spoken or written product.

Methods of Testing

Testing was implemented online using a Moodle Content Management System (CMS) on university servers. From the objectives, pre-test and post-test items were created for each subtitled course. For the most part, the pre-test and post-test items per subtitled course were identical, the exceptions being some of the subtitled course tests contained items which required the students to input their opinions as short answer responses. Test items and answer choices within the items were randomized either automatically through the CMS or manually, to create randomized test sections on particular tests.

Students were directed via a paper handout written in Japanese with instructions to access the online tests during the first three weeks of the course (the pre-test period) which coincides with the student registration period for classes. The post-test period spanned the last two weeks of the courses. The majority of tests were completed outside of class as computer rooms were not available during the majority of course time slots; however, when facilities were available, some full-time teachers allowed time for tests to be taken during class time. The time limit for each subtitled course test was set between 30 and 40 minutes and each test was composed of 35 to 45 items. The majority of items were multiple-choice; however, cloze items and short answer writing items were used on some subtitled course tests. Completing all items on both of the pre-course and the post-course tests allowed students to receive 10% toward their course grades.

Methods of Materials and Teaching

The selection of materials and teaching to the course objectives were the responsibilities of the individual teachers. Some teachers taught from commercial textbooks while others used teacher created materials. Teachers were responsible for 90% of the students' course grades, including but not limited to in-class assignments, homework, quizzes, tests, presentations, effort, and participation.

Methods of Course Evaluation

To gather feedback in the first and second semesters of 2009, a voluntarily online post-course survey was available for students to complete before or after the tests. However, in 2010 both pre-course and post-course surveys were made available online to students for voluntary submission after completion of the online pre-tests and post-tests. The pre-course survey focused mainly on the students' interests, such as their reasons for taking the course and what extra-curricular English language activities they may be interested in. The post-course survey focused on course feedback. Post-course teacher surveys, which consisted of questions which mirrored those on the students' post-course surveys, were distributed and collected by postal mail and email at the conclusion of each course.

Overall EC course evaluation was done after each semester. The evaluation phase took into account feedback from tests, student surveys, teacher surveys, and other communications with students and teachers. After each semester, item analyses of the online tests were done for each subtitled course test to improve test items. Also, after each semester, feedback from the student and teacher surveys and direct feedback from teachers and students were analyzed. Using the results from the online achievement tests and the student and teacher surveys, the objectives were adjusted annually.

Results

Results from Online Tests

Between ECI 2009 and ECII 2010, student achievement for each course ranged between -1% and 24% (Table 2). Please note that the subtitled courses were implemented over a two-year period with all courses being offered in ECII 2010. The ECII 2010 CA2 results were incalculable due to sample size (n = 2). In 2011 online testing and surveys were suspended due to rolling blackouts and to conserve energy after the Great Tohoku Earthquake, yet the teachers still taught according to the guidelines.

Table 2. ***Average Student Achievement***

	ECI 2009	ECII 2009	ECI 2010	ECII 2010
TA1	1% (n = 243)	6% (n = 64)	3% (n = 169)	3% (n = 91)
TA2	-1% (n = 48)	12% (n = 94)	24% (n = 49)	18% (n = 14)
PS1	5% (n = 151)	1% (n = 86)	4% (n = 100)	5% (n = 91)
PS2	-	5% (n = 38)	12% (n = 15)	9% (n = 33)
BU1	-	-	7% (n = 151)	5% (n = 62)
BU2	-	-	-	5% (n = 51)
CA1	-	-	0% (n = 21)	4% (n = 17)
CA2	-	-	-	NA

Results from Surveys regarding Expectations, Enjoyment and Learning

Regarding expectations, student and teacher survey data (Figure 1) indicated that the majority of students felt that they were achieving the spoken and written expectations, except in the first semester of 2009. Respectively, from 2009 in each of the four semesters, 49%, 71%, 62%, and 72% of the students who completed the survey thought that they had achieved the writing expectations for the course. Regarding the speaking expectations, in semester order from ECI 2009, 80%, 89%, 86%, and 90% of the students surveyed perceived that they had achieved the expectations. Teacher perception of achieving the expectations advanced each semester in regard to writing, 85%, 86%, 94%, and 100%. The speaking expectations were perceived to be achieved 94% in ECI 2009 and 100% in all semesters following by all teachers. These data indicated that as the EC courses were further developed, both students and teachers more often felt that both the writing and speaking expectations were achieved.

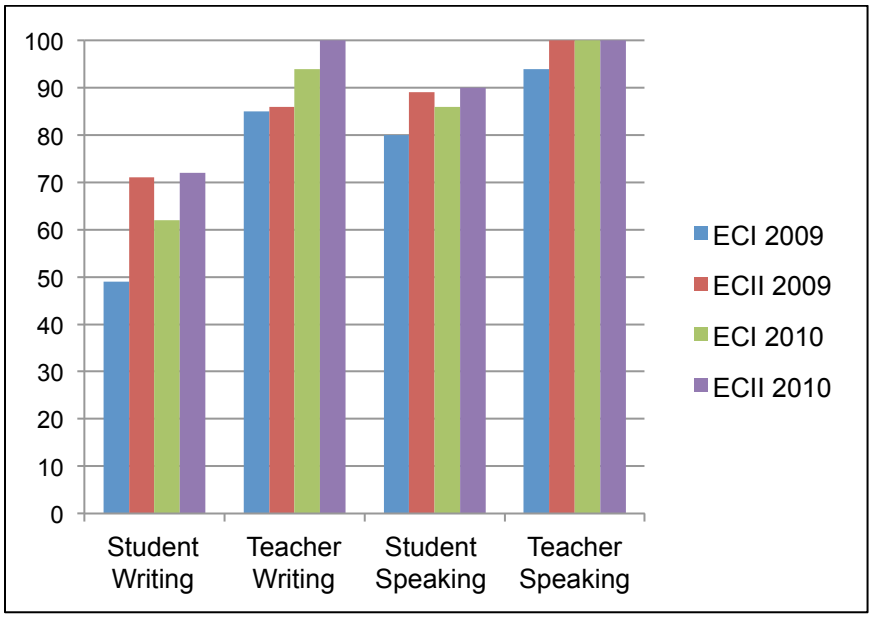

Figure 1. **Perceptions of speaking and writing expectation achievement**

Regarding enjoyment and learning, data from 2009 and 2010 pre-course and post-course student surveys gave some indication of student satisfaction. Figures 2 and 3 compare students' perceptions regarding pleasure and learning over four semesters. It is important to note that the majority of students, approximately 85% each semester, consistently enjoyed the courses, 10 – 15% were undecided, and 1 – 4% disliked the courses. While enjoyment of the courses has been steady at 85%, a noteworthy tread in the data is that each semester between ECI 2009 and ECII 2010 the students who learned "a little" has decreased, and the students who indicated they learned "a lot" has steadily increased.

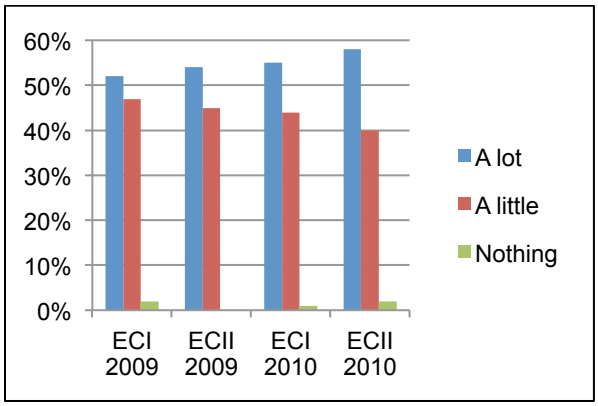

Figure 2. **Student perceptions on learning**

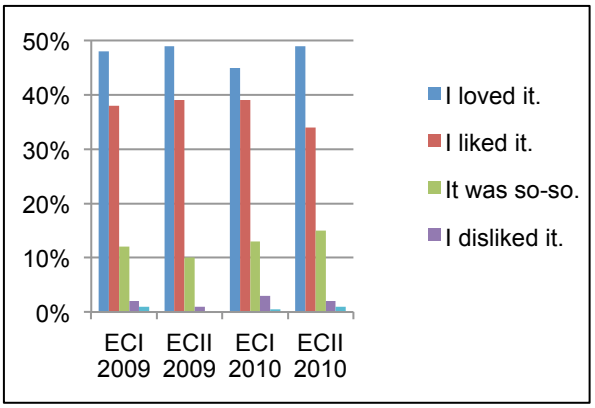

Figure 3. **Student perceptions on enjoyment**

Results regarding Participation and Overall EC Evaluation

Between 650 and 850 students register for EC each semester, but for a more accurate picture of enrollment, test and survey participation was summarized (Table 3). The number of students who were given credit for completing the online tests and the number of voluntary surveys completed are compared with the number of students who attended class in the third week, which was the last week to register for the course. The third week attendance data most accurately reflects the actual number of enrolled students

who attended the courses as official registration statistics tend to run high and post-test completion numbers tend to run low. Please note that ECII attendance is lower than ECI attendance due to schedule conflicts with required second semester major-specific courses.

Table 3. ***Test and Survey Completion Compared with Attendance***

Semester & Test	3rd Week Attendance	Tests Completed	Surveys Completed	Credit Given
ECI 2009 Pre	768	739	345	456
ECI 2009 Post		598		
ECII 2009 Pre	615	611	237	326
ECII 2009 Post		513		
ECI 2010 Pre	739	700	468	519
ECI 2010 Post		622	430	
ECII 2010 Pre	561	567	375	384
ECII 2010 Post		512	379	

Illustrated in Figures 4 and 5, the data from Table 3 show that from ECI 2009 to ECII 2010 completion of both pre-course and post-course tests increased from 45% to nearly 70% and completion of the voluntary surveys increased approximately 10% from below 60% to approximately 70%.

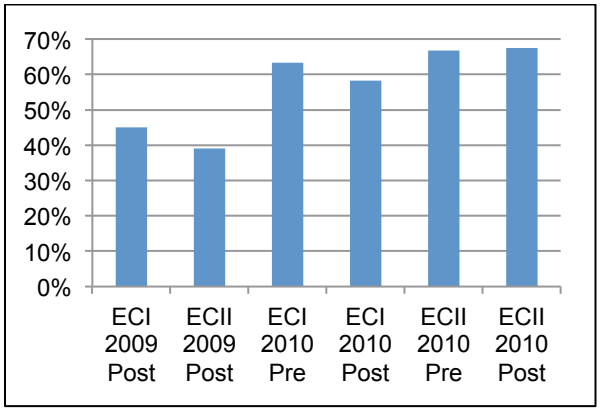

Figure 4. **Survey participation**

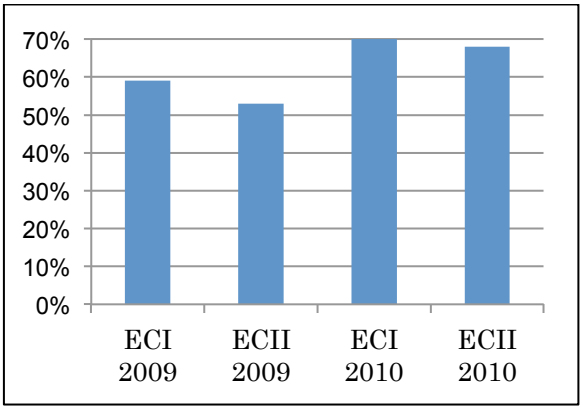

Figure 5. **Test participation**

Discussion of Course Evaluation and Improvements

Since its inception in 2009, EC has been evaluated and updated after each semester. The results from the online tests, online student surveys, teacher surveys, communications with students and teachers were all used to evaluate EC and update aspects of the courses. To improve the guidelines, the objectives were better defined each year. General skills objectives for note-taking and paragraph writing were added. Also, an objective for email writing was implemented as an early-on needs analysis regarding technology use indicated that students were using mobile technology and had access to personal computers (Harrison, 2009), yet teachers noted email communications with students were impolite and not easily understood. For results regarding email learning, see Harrison & Vanbaelen, 2011. Also, the expected amount of spoken and written output for Level 1 and Level 2 courses was raised. Open email communication with students and teachers has been promoted, and annual faculty development sessions have been held. The above improvements are believed to have raised achievement and participation.

Conclusion

Overall, Brown's approach is flexible and effective in promoting education. Results on achievement tests and feedback from surveys were increasingly positive through 2009 and 2010. The students' and teachers' perceptions of spoken and written output goal achievement increased each semester, with the students' perception of goal achievement rising to 72% for writing and 90% for speaking. Achievement test results for 2010 showed 3% to 24% achievement for subtitled courses. Data also indicated more than 85% of learners enjoyed the courses and 98% of the students perceived they were learning with the percentage of those who felt they were learning "a lot" on the rise.

References

Brown, J. D. (1995). *The Elements of Language Curriculum: A Systematic Approach to Program Design*. Boston: Heinle & Heinle.

Harrison, J. (2009). A needs analysis for incorporation of technology into English courses. *Nihon University, College of Science and Technology Bulletin of the Department of General Education*, College of Science and Technology, Nihon University, 86, 23-32.

Harrison, J., & Vanbaelen, R. (2011). Learning and retention of English email writing skills by students at an engineering university in Japan. *Professional Communication Conference (IPCC), 2011 IEEE International*, 1-7. doi: 10.1109/IPCC.2011.6087197

Nihon University, College of Science and Technology, Office of Educational Affairs. (2009-2011). Personal communications by Jonathan Harrison.

Software Corner:

RKWard: IRT analyses and person scoring with ltm

Aaron Olaf Batty
abatty@sfc.keio.ac.jp
Keio University
Lancaster University

In *SRB 16*(2), I introduced the ever-improving, free and open-source RKWard GUI for the free and open-source R statistical package (Batty, 2012). In this issue, I will detail how to conduct IRT analyses from RKward's drop-down interface using `ltm`, a flexible R package for unidimensional item response models. This allows users to get many of the important features of commercial software packages such as Xcalibur and IRTPro for free, and with a user-friendly interface.

Updates

Since that article, the RKWard development team has been hard at work improving the package for Windows and Macintosh users, and the following changes have since been made to installation practices:

- The RKWard project page has been much improved, and can be found at the following address:

  http://rkward.sourceforge.net/

- On Windows, the installation process is now simply a matter of decompressing a .zip file and running the .exe application within the created folder. The folder can be saved in your Program Files or Program Files (x86) folder, with a shortcut on the Start menu/screen if you so desire, but it does not necessarily need to be. You can store it anywhere and run it from anywhere on your hard disk.

- On the Mac, the installation method is to first install the standard R package, then install the RKWard GUI. When downloading the RKWard package, make note of the file name, as it specifies what version of R you should install. Previously, RKWard installed its own instance of R, but this resulted in a bloated installation that was difficult to maintain. R can be downloaded from the following mirror in Japan:

  http://cran.md.tsukuba.ac.jp/

In the year since I first introduced RKWard in these pages, it has become my main statistical package. It features SPSS' ease of use and the power, flexibility, and price of R. Despite the fact that I have a personal copy of SPSS, and site licenses through both of my institutions, IBM's licensing controls are simply too onerous to justify using it when there is a free alternative.

In this issue, I will explain how to use the `ltm` IRT package via RKWard to fit Rasch and other IRT models; plot item characteristic curves; check item, person, and model-data fits; compare models for selection; and output person ability estimates.

Notes:

- *I will be demonstrating with the Macintosh version of RKWard, but the Windows version is identical, since the software itself runs in a KDE environment. The only real difference is that the*

menubar for the Windows version is in the RKWard window, whereas on the Mac it is at the top of the screen.

- *Sometimes it is necessary to click the "Refresh Output" button on the output display for the results of an analysis to appear.*
- *This is not intended to be a comprehensive guide to* ltm*. Much more information is available in the package reference manual (Rizopoulos, 2013).*

The ltm package

The ltm—which stands for "latent trait models"—package is a set of functions for R, developed by Dimitris Rizopoulos (Rizopoulos, 2006). The package offers a wide array of testing and IRT functions, including 1PL, 2PL and 3PL dichotomous models, Rasch and Generalized Partial Credit Models and the Linear Logistic Test Model, which is similar to Linacre's many-facet Rasch model. Many of these models can be run via RKWard's GUI, and parameters can further be constrained to fixed values specified by the user via the command line. For a full overview of functions, see the project page at the URL below, or in the reference manual (Rizopoulos, 2013):

http://rwiki.sciviews.org/doku.php?id=packages:cran:ltm

The ltm package is not installed by default with RKWard; however, the first time you attempt to use it, RKWard will walk you through choosing a download mirror (there are three in Japan), after which you merely have to click "Apply" and then "OK." On the first use, or the first use with a workspace, some messages will display at the bottom of the output telling you that it has been installed and/or loaded for the analysis.

Preparing your data

As explained in my last column, data need to be imported to RKWard in either delimited text or SPSS .sav format. Your data for IRT analyses must conform to the following:

- Numeric responses, either binary for dichotomous datasets, or integer data for polytomous.
- No person IDs. Item responses only.
- (Optional) Item names in the header row, which can be imported as variable names.

As a general rule, when importing datasets to RKWard, I prefer to name them "[descriptive name].data", e.g., "Listening.data", rather than keeping the default "my.csv.data", etc., that RKWard enters automatically. Data tables can also be renamed in the Workspace pane by right-clicking and selecting "Rename."

Fitting IRT models

A wide array of IRT models, both dichotomous and polytomous, can be fitted; the process for each, however, is the same.

1. Go to the following menu entry:

   Analysis → Item Response Theory → Dichotomous data → *[choose the model you want]*

2. A window such as that shown in Figure 1 appears. Select your test data from the list on the left and click the green triangle to select it for analysis.

3. (Optional, but recommended) Change the name of the output at the bottom to reflect the dataset. This is especially important if you have multiple datasets in your workspace. See Figure 1 for an example.
4. Click "Submit."

The parameter estimates will be displayed in a table in the output. They can then be exported or copy/pasted into another program (e.g., Excel) for saving. In addition, an item estimates object will be added to the workspace pane, which will be used for further analyses, some of which will be explained later in this article.

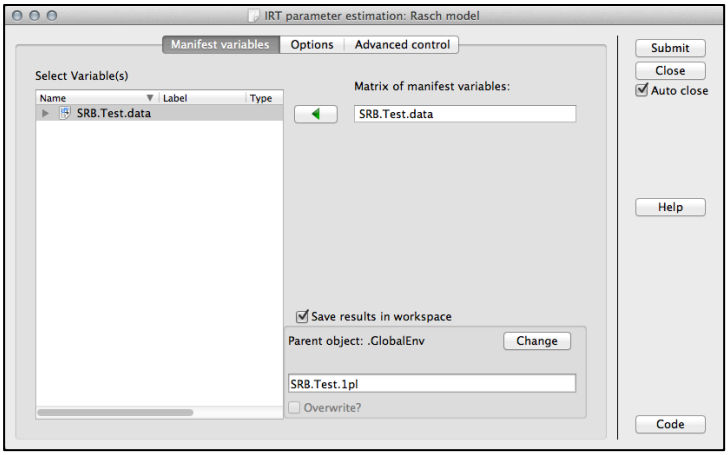

Figure 1. **Selecting the data for IRT model fitting and defining the output name**

Fitting a traditional Rasch model

It is important to note that the default Rasch model fitted by `ltm` is perhaps better described as a one-parameter IRT (1PL) model, as the discrimination parameter (or "slope") is estimated separately for each data set, rather than assumed to be 1 as it is in programs such as Winsteps (Linacre, 2012). Such a model retains the item invariance that is the main "selling point" of the traditional Rasch model, but also allows better fit with test data which have higher or lower average item discrimination, which can allow for a more realistic picture of the test information function (test reliability by ability levels). For a graphical representation of the difference, see Figure 2.

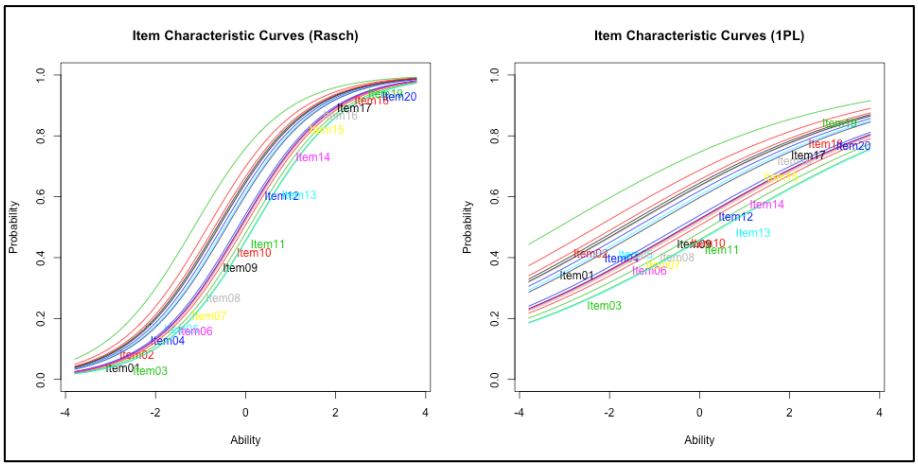

Figure 2. Comparison of Rasch and 1PL item characteristic curves of the same data

If you would prefer to fit a constrained Rasch model, there are two extra steps.

Preparing a constraint matrix

To constrain the model with a discrimination parameter of 1, an initial step is necessary (see Figure 3).

1. Click the "R Console" button in the lower left of the RKWard window.
2. Enter the following R command:

   ```
   [Some descriptive name] <- cbind([number of items in your test, plus one], 1)
   ```

   e.g., for a twenty-item test:

   ```
   TwentyItemConstraint <- cbind(21, 1)
   ```

3. Hit enter/return.

This will create a new object in your Workspace pane (click the vertical "Workspace" button in the upper left-hand side of the RKWard window to view it) with whatever name you've assigned it ("TwentyItemConstraint" in my example), which is a matrix for setting the discrimination estimate to 1.

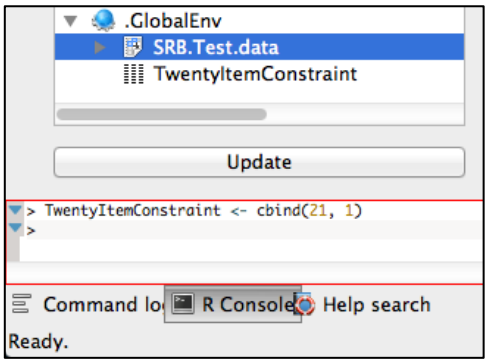

Figure 3. **R command for creating a constraint matrix, and its display in the Workspace pane**

Fitting the Rasch model

With the constraint matrix created, fitting the Rasch model is simple:

1. Go to the following menu entry:

   Analysis → Item Response Theory → Dichotomous data → Rasch model fit

2. Follow steps 2 and 3 from the "Fitting IRT Models" section above.
3. Click the "Options" tab. The window shown in Figure 4 will appear.
4. Select your constraint matrix and click the green triangle as shown in Figure 4.
5. Click "Submit."

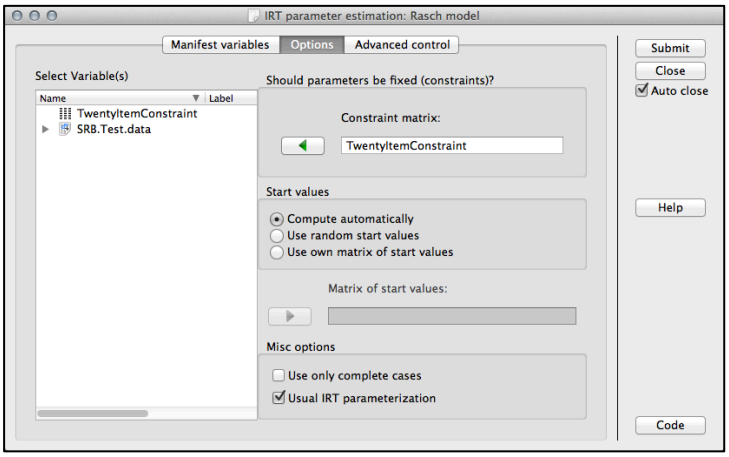

Figure 4. **Selecting the constraint matrix to constraint discrimination estimates to 1.**

Plotting IRT item characteristic curves (ICCs)

One of the most intuitive methods of interpreting item behavior is by viewing ICC graphs (see Figure 2 for an example). `Ltm` can create these *once you have already estimated a model and the estimate output object is available in your workspace pane.*

ICC graph settings are accessible from the following menu:

Plots → Item Response Theory → *[choose the model you want]*

Once again, dichotomous and polytomous data each have their own submenu. Choose the appropriate plot for the model you have already estimated and follow the instructions below:

1. Select the appropriate ICC plot type from the submenus. The window shown in Figure 5 opens.
2. Select the estimates object created when you estimated your IRT model from the list on the left and click the green triangle.
3. (Optional) Click the "Plot type" tab and select the kind of plot you want. You can also specify subsets of items here. The default is to create a graph with all the ICCs displayed, but other options are available as well.
4. (Optional) Further options are available in the "Options" tab, especially by pressing the "Generic plot options" button, where you can define the title, subtitle, and other such plot characteristics.
5. Click "Submit."

Your selected plot will display in the output. Consult my previous article for instructions on how to export these.

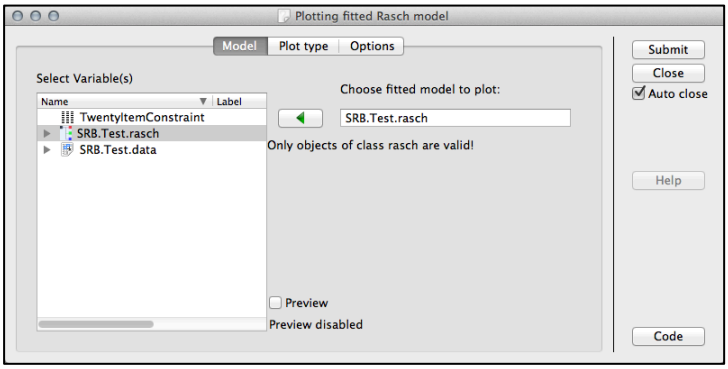

Figure 5. **Selecting the model output for plotting.**

Checking and comparing fit

The ltm package provides a wide variety of fit statistics, both for items and persons, and for entire models. For items, persons, and Rasch models, fit can be determined via chi-square tests. It is also possible to compare fits between two nested models—for example, a 2PL model versus a Rasch model, in order to determine if the additional parameter substantially improves fit.

Item and person fit

Item and person fits statistics are provided as chi-square statistics comparing the observed responses to those predicted by the model. These tests are available from the following menu:

Analysis → Item Response Theory → Tests → *[select either item-fit or person-fit statistics]*

The process is as follows:

1. Select the estimate output from your model estimation (e.g., "SRB.Test.rasch") and click the button with the green triangle.
2. Click "Submit."

Fit statistics are provided as a table in the output, listing item names, chi-square statistics, and p-values. A significant chi-square statistic indicates that the item or person responses differ significantly from the model-expected responses, i.e., the item or person *misfits* the model.

Model-data fit for a Rasch/1PL model estimation

For one-parameter models, a chi-square test for the fit of the entire model is available from the following menu:

Analysis → Item Response Theory → Tests → Goodness of fit (Rasch)

The process is the same as the other analyses discussed in this article. The estimation output object is selected, and the button with the green triangle is clicked, followed by the "Submit" button.

NOTE: It may take a long time to complete the analysis. Be sure to look at the R engine status indicator in the lower right of the RKWard window. If it is red, the engine is still working.

The result of the chi-square test is displayed in the output. Once again, a significant p-value indicates that the model-expected and observed values are significantly different, hence the fit is *poor*. A non-significant chi-square indicates good model-data fit.

Comparing model-data fit

When choosing an IRT model, it is important to choose the most parsimonious model that fits the data (Kang & Cohen, 2007). Item response models with additional parameters will almost always have an at least negligibly better fit than models with fewer parameters, but the more complex model may simply overfit the data and provide information that will not be generalizable to another sample. Luckily, `ltm` offers two methods of model comparison to aid in selection.

Rasch/1PL models

If you want to compare two one-parameter (i.e., Rasch or 1PL) models to determine which has superior fit with the data, the easiest method is to simply run the Rasch goodness of fit test discussed above for each and compare the chi-square fit statistics.

`Anova` *command for use with other models*

`Ltm` includes a convenient command for comparing the fits of two nested models, but it does not, unfortunately, have a graphical interface in RKWard. Luckily, though, RKWard allows the user to use a standard R console as well. To use the `ltm anova` command, do the following:

1. Click the "R Console" button on the bottom left of the RKWard window (see Figure 6).
2. Type the following at the R prompt:

   ```
   anova([the output object for the simpler model], [the object for the more complex model])
   ```

   e.g., to compare a Rasch model to a 2PL model (see Figure 6):

   ```
   anova(SRB.Test.rasch, SRB.Test.2pl)
   ```

3. Hit enter/return.

The result is displayed in the console (see Figure 6). It will also eventually appear in the output pane when another command is issued through the GUI.

The `anova` command compares the first object, the null, to the second, the alternative, using a number of statistics for nested models, including the Akaike Information Criterion (AIC) the Bayesian Information Criterion (BIC) and a Likelihood Ratio Test (For more information on model fit comparison, see Kang & Cohen, 2007.). In contrast to the chi-square statistics discussed above a significant p-value for the likelihood ratio test indicates that the alternative model fit is *superior* to that of the more parsimonious model.

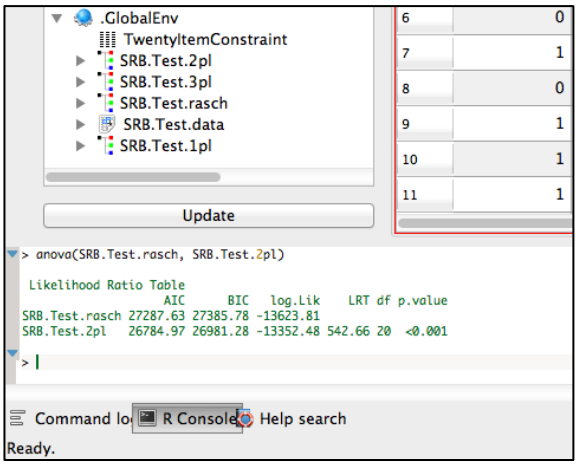

Figure 6. `ltm anova` command and output in the R Console.

Estimating and exporting person ability estimates

`ltm` includes several methods of estimating person ability (see Rizopoulos, 2006, 2013 for details), but they are all accessed with the `factor.score` command. In this section I will use the default empirical Bayes estimation method. Unfortunately, there is no GUI for person scoring in RKWard as of this writing.

Generating ability estimates

To estimate ability scores, do the following:

1. In the R Console at the bottom of the RKWard window, type the following (see Figure 7):

   ```
   [name of the person score object you want] <- factor.scores([name of the model estimation output you want to score])
   ```

   e.g.:

   ```
   RaschScores <- factor.scores(SRB.Test.rasch)
   ```

2. A new object is added to the workspace pane (called “RaschScores” in this example). Click the expansion triangle next to it and locate the data table inside called “score.dat”.

3. Double-click score.dat to open it in the main RKWard pane. The person measures and their SEs are listed in the last two columns of data, labeled “z1” and “se.z1”. See Figure 7 for an example.

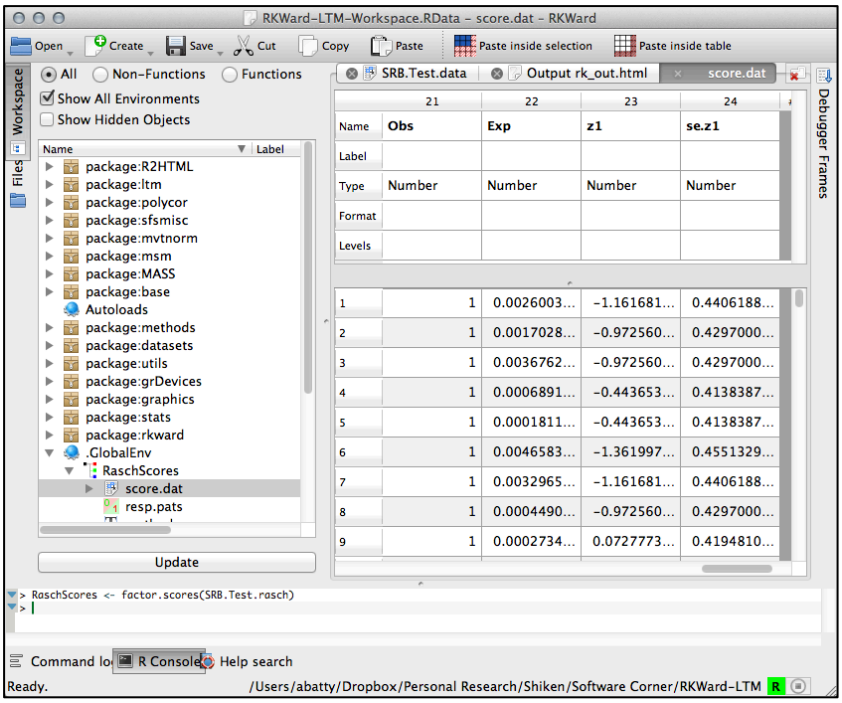

Figure 7. **`factor.score` command, output object, and score data.**

Exporting ability estimates

Although I discussed exporting *outputs* in my last RKWard article, to be able to use the ability measures elsewhere (e.g., in Excel), they need to be exported. To export tabular data to a file:

1. Go to the following menu:

   File → Export → Export tabular data

2. The window shown in Figure 8 opens. Select the score.dat table and click the button with the green triangle.
3. Under "Name of the file," click the folder button and navigate to where you would like to save the file, give it a name, and click "OK."
4. (Optional) Because the first column simply contains the case numbers with no heading, the header row will be off by one column if you export with the default settings. To remedy this, click on the "Columns and Rows" tab and select "No names" under "Name of rows" in the middle of the window.
5. Click "Submit."

The file that is created is simply a tab-delimited text file that can be opened in Excel, etc. It may be necessary to start your chosen software first and direct it to open the file, or to right-click the file and select which application to use. The delimiter can be changed (e.g., to a comma) under the "Options" tab before clicking "Submit."

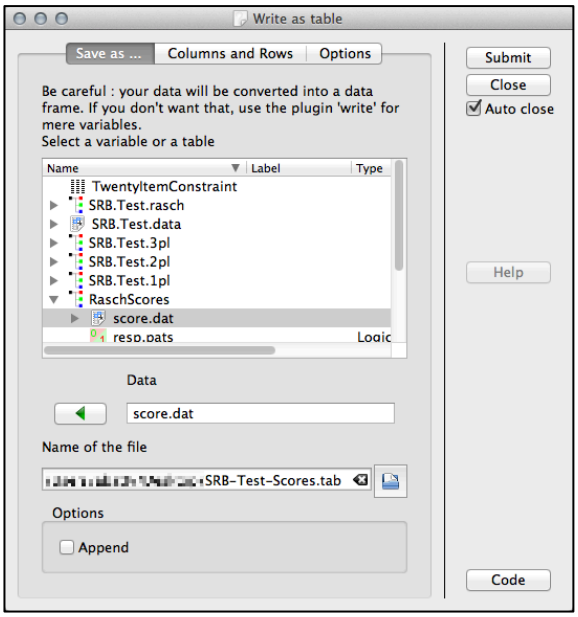

Figure 8. **Export dialog window for tabular data**

Conclusion

I have only introduced some of the basic functions available in `ltm` in RKWard, and even more functionality is available on the command line. Although commercial IRT software is often easier to use, and sometimes creates more versatile plots and outputs, you may find that `ltm`, especially when paired with the RKWard GUI, offers features that meet your testing or research needs at no cost whatsoever.

References

Batty, A. O. (2012). Software corner: RKWard: Installation and use. *Shiken Research Bulletin*, *16*(2), 21–29. Retrieved from http://teval.jalt.org/sites/teval.jalt.org/files/SRB-16-2-Batty-SoftCorner.pdf

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331–358. doi:10.1177/0146621606292213

Linacre, J. M. (2012). *Winsteps*®. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jwalkonline.org/docs/Grad%20Classes/Spring%2008/modmeas/homework/ltm.pdf

Rizopoulos, D. (2013, February 15). ltm reference manual. Retrieved from http://cran.r-project.org/web/packages/ltm/ltm.pdf

Rasch Measurement in Language Education Part 8:

Rasch measurement and inter-rater reliability

James Sick
American Language Institute, Tokyo Center
New York University School of Continuing and Professional Studies

The previous installment of this series dealt with how many-facet Rasch analysis (MFRA) can be used to adjust for differences in rater severity when measures are constructed from subjective judgments. This installment addresses the related issue of rater agreement and inter-rater reliability from the perspective of classical test theory (CTT) and Rasch measurement theory (RMT).

Question:

I recently completed a many-facet Rasch analysis using Facets on of a set of essays marked by 8 teacher raters. Each essay was marked by two raters on a nine-point scale. If ratings differed by more than two points, a third rater was asked to arbitrate. I would like to know the inter-rater reliability of this assessment. Facets Table 7, the Raters Measurement report, indicates a reliability of .99 but states that this is NOT inter-rater reliability. If not inter-rater reliability, what is it? Where in Facets can I find inter-rater reliability? Table 7 also has a column indicating percentages of "Exact Agreement," observed and expected, for each rater. I assume these refer to rater agreement statistics, which I understand, but expected agreements vary amongst raters, ranging from 22 to 45 percent. Why are some raters expected to agree more than others, and for that matter, why are they all not expected to agree 100 percent?

Answer:

In classical test theory, the reliability of tests requiring subjective judgments, such as essays or speaking performances, is generally assessed using agreement ratios or inter-rater reliability. There is no universally agreed index of inter-rater reliability, however. A correlation coefficient can be used to assess the degree to which two raters agree in their rankings. When more than two raters are involved, an average correlation or a correlation adjusted with the Spearman-Brown formula may be used as an index of overall agreement in rankings (see Brown, 1996). I say "rankings" because the correlational approach can obscure systematic differences in rater severity. For example, if one rater awards scores of [5, 4, 3, 2] while another awards scores of [4, 3, 2, 1], the raters will be perfectly correlated, even though they do not agree in their assessments. Another approach is to calculate the proportion of exact or nearly exact agreements amongst raters. Agreement ratios are easy to understand and have practical significance when disagreements above a certain threshold prompt an additional rating. Cohen's Kappa and Fleiss' Kappa are refinements of the agreement approach (see Brown, 2012). Both derive reliability indices from the percentage of exact agreements, adjusting for the probability that agreements can occur by chance. All of these approaches share the common aim of separating the variance due to examinee performance (the true score) from the variance due to the vagaries of subjective judgment (the error).

Before answering your questions, it will be helpful to discuss some philosophical differences between CTT and RMT regarding subjective assessment. In CTT, an essay score is based on rating descriptors and is considered an attribute of the essay. The goal of the raters is to apply the rating rubric with machinelike consistency. Rater disagreement is seen as an indication that one or both raters have not assigned the most appropriate score, and is thus regarded as undesirable. In the CTT approach, test quality is pursued by minimizing disagreements through calibration sessions, fine-tuning the descriptors,

and arbitration when large disagreements are identified. Test reliability is estimated by inferring, from their level of agreement, the degree to which raters are identifying the true score.

In contrast, raters in a many-facet Rasch analysis are viewed as independent experts who will sometimes disagree in their assessments. Although calibration sessions and descriptors are considered useful for achieving a shared understanding of the construct, individual essays are regarded as inherently too complex to be consistently matched to a set of descriptors. From the Rasch perspective, essay evaluation is better accomplished by engaging experts with multiple perspectives to respond to the unique characteristics of each individual essay. Scores awarded by independent raters, however, cannot be used "as is" because they are the product of both rater and examinee characteristics. Instead, initial scores are an intermediate step: data that can be used to construct measures of the underlying attribute, in this case English writing ability, that produced the variation in the essays.

In RMT, raters are hypothesized to possess a psychological trait that we label "severity" which leads them to systematically award higher or lower scores than their co-raters. Variation in severity arises from differences in personality, culture, and experience, but most likely reflects the fact that some people are adept at spotting flaws, while others tend to focus on strengths. Both perspectives are valuable, provided that severity is taken into account when constructing measures or making decisions. A many-facets Rasch analysis uses rater disagreements to estimate severity, and adjusts examinee measures accordingly. Rater severity is expected to influence the initial scores in a probabilistic manner consistent with the Rasch model. This is not a given, however. The analyst must verify that the raters fit the Rasch model as part of the process of validating the assessment.

To better address your questions, I've included a Facets raters measurement report from an analysis similar to the one you described (Table 1). This table reports rater severity and fit statistics for ten raters who provided two ratings for 396 essays. These raters varied considerably in severity, ranging from a high of 2.40 logits to a low of -2.60 logits. The reliability index at the bottom of the table indicates the reliability of the severity measures. That is, the degree to which these severity measures would be reproduced if the raters evaluated a similar sample of essays. Reliability is high because there is a lot of variance in rater severity and a large sample of shared ratings (agreement opportunities) from which to estimate it. If you are conducting a study of rater behavior, this reliability will be of interest to you. If your chief interest is examinee performance, however, it is not particularly relevant. Nevertheless, let me point out that a low reliability in the rater measurement report indicates a high level of rater agreement. If raters agreed 100 percent, there would be no variance in severity and the reliability would be zero. If your situation requires high rater agreement for reasons of face validity, a low reliability in the rater report is desirable.

Table 1. ***Raters Measurement Report (Facets Table 7.3.1)***

Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
2.40	.19	.67	-2.1	.66	-2.2	20.0	27.7	2	Hayward
2.20	.20	.65	-2.2	.63	-2.2	21.1	24.8	6	Garland
2.08	.20	.79	-1.2	.75	-1.5	30.9	43.8	10	Brando
.70	.20	.93	-.3	.99	.0	23.0	29.7	3	Cruz
-.41	.21	.50	-3.4	.52	-3.1	37.3	35.8	5	Leigh
-.84	.18	.95	-.2	.94	-.3	29.8	42.6	1	Grable
-1.04	.19	1.03	.2	1.08	.5	17.6	30.3	4	Temple
-1.07	.17	.97	-.1	1.02	.1	23.7	33.7	8	Monroe
-1.43	.19	1.61	3.0	1.53	2.7	26.3	41.8	9	Rogers
-2.60	.20	1.42	2.1	1.47	2.3	19.0	29.2	7	Davis
.00	.19	.95	-.4	.96	-.4				Mean (Count: 10)
1.65	.01	.33	1.9	.32	1.9				S.D. (Population)
1.74	.01	.34	2.0	.34	2.0				S.D. (Sample)

Model, Sample: Separation 8.95 Reliability (not inter-rater) .99
Inter-Rater agreement opportunities: 396
Exact agreements: 97 = 24.5%
Expected: 134.9 = 34.1%

As for "where can I find inter-rater reliability in Facets," that statistic belongs to the realm of CTT, and Facets does not report it. The Rasch equivalent would be the reliability reported in the Examinee Measurement Report, Facets table 7.1.1. This index answers the question "how likely would these measures be reproduced if the examinees produced another set of essays that were evaluated by a similar sample of raters?" I have not included the examinee report here, but for the analysis above, examinee reliability was .91. Let me emphasize that this figure applies to the Rasch measures, as opposed to the raw scores. The rater measurement report indicates that the raw scores were highly influenced by rater differences and are not very dependable as indicators of examinee performance. If inter-rater reliability were calculated, I would expect it to be rather low.

Column 3 of Table 1 reports the ratios of observed and expected exact agreements. The reason why expected agreements are not 100 percent should now, I hope, be clear. The reason expected agreement varies amongst raters is that the probability of agreement is dependent on the relative severity of the co-raters. Hayward and Garland, for example are similar in severity and would be expected to agree with each other frequently. Hayward and Davis, on the other hand, are nearly 5 logits apart and are expected to agree only rarely. Expected agreements for individual raters depend not only on their own severity, but also on the severity of the raters they were paired with.

The validity of the Rasch measures requires that rater severity be consistent in a manner that fits the probabilistic expectations of the Rasch model. Sizeable discrepancies between observed and expected agreements are an indication that raters are not consistently severe or lenient. This is also expressed in the infit and outfit statistics shown in Column 2. Rogers, for example, has an infit mean square of 1.61, indicating that his behavior does not fit the Rasch model very well. With a severity measure of -1.43, he is overall a lenient rater but frequently awards scores more or less severe than expected. This is confirmed in Column 3 where we see that his observed agreements are much less than expected. Leigh has an infit mean square of .50 and more observed agreements than expected. This is an indication that she is not behaving like an independent expert. She may be avoiding disagreement by keeping her scores in a safe middle range, or she may be colluding on her assessments with another rater. Comparing the total exact agreements to the total expected agreements, as shown at the bottom of Table 1, provides a global assessment of rater fit. In this example, observed agreements were less than expected, indicating that raters generally acted independently, but were not always consistent in severity.

To summarize, rater agreement is viewed differently in CTT and RMT. In CTT, disagreement amongst raters is seen as a source of error that must be minimized. Inter-rater reliability indices reflect the CTT approach and are generally not reported by Rasch programs such as Facets. In RMT, rater disagreement is seen as an indication that raters are behaving independently, bringing multiple perspectives and differing expertise to their assessments. In RMT, rater disagreement is used to estimate rater severity, which can then be employed to adjust performance measures for the detrimental effects of rater independence. For adjustments to be valid, however, it must be verified that raters are consistent in severity in a manner that conforms to the expectations of the Rasch model. Rater fit statistics and observed to expected agreement ratios can be used to assess rater fit. If the data reasonably approximate the Rasch model, the reliability reported in the examinee measurement report is the nearest equivalent, in purpose and substance, to inter-rater reliability as employed in CTT.

References

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

Brown, J. D. (2012). Statistics Corner: How do we calculate rater/coder agreement and Cohen's Kappa? *Shiken Research Bulletin, 16*(2), 30-36. Retrieved Dec 2, 2013 from http://teval.jalt.org/sites/teval.jalt.org/files/SRB-16-2-Brown-StatCorner.pdf

Statistics Corner:

Solutions to problems teachers have with classroom testing

James Dean Brown
University of Hawai'i at Mānoa

Question:

I sometimes feel like I must be making lots of mistakes when I write tests for my students. What worries me most is that I may be wasting my time and theirs because I don't know what I am doing. Can you help me by explaining common mistakes that teachers make when they design tests and how to avoid them?

Answer:

The problems that test designers have when writing and developing standardized tests (norm-referenced tests) are discussed in many language testing books. However, the problems that teachers have in implementing classroom tests (criterion-referenced tests) are rarely covered. Yet surely, testing occurs more often in language classrooms than in standardized language testing settings. So I will be happy to address the classroom testing problems that teachers face and offer solutions to those problems—at least to the best of my ability. I will do so in three sections about problems that teacher may have in test writing practices, test development practices, and test validation practices.

Test Writing Practices

In test writing practices, teachers sometimes have problems with: creating good quality test items, organizing those items in the test, and providing clear headings and directions.

Create good quality test items. The biggest single *problem* that most teachers have with tests is the tendency to treat tests as an afterthought, waiting until the last possible moment to write a test for the next day. This habit leaves teachers with too little time to create good quality items. In many cases, I suspect that this tendency is caused by lack of training in item writing, training that, if nothing else, would teach them that writing good test items takes time.

Clearly, the *solution* to this problem is to get ahold of a good book on language testing and read up on what good quality items are and how to write them (see e.g., Brown, 2005, pp. 41-65; Brown & Hudson, 2002, pp. 56-100; or Carr, 2011, pp. 25-45, 63-101). Then when it is time to write a test, make sure to allot enough time for writing good quality test items by starting early. These strategies will pay off handsomely because a carefully written test will always be better than a shoddily written one.

Organize the items. The *problem* here is that tests sometimes seem like a disorganized hodge-podge. Any test will be clearer to the students and easier for them to negotiate if the items are clearly organized into sections that make up the whole test. Teachers naturally try to organize their tests, but this can always be done better.

The *solution* is to follow at least three basic principles: (a) group items that are testing the same language point together, (b) collect items of the same format (e.g., multiple-choice, true-false, matching, writing tasks, etc.) together, and (c) group items based on reading or listening passages together with the

passages they are based on. Unfortunately, these three principles are sometimes in conflict. For instance, it would be reasonable to have a test with say three reading passages; each reading passage might have one multiple-choice main-idea item, one fact item, one vocabulary item, and one inference item, and each passage might be followed by an open-ended critical-thinking item that students must answer in writing. Clearly, such a test would be following principle (c) above but not (a) and (b). Another section on the same test might group multiple-choice questions together with five for articles, five for prepositions, five for copula, and so forth. That would be following principles (a) and (b) but not (c). I stand by these three principles, but they are not hard and fast rules, and they may not all apply at the same time. Common sense should guide which of the three need to be applied and in what combinations.

Provide clear headings and directions. The *problem* is that even when a test is well-organized, the students may not understand that organization, or worse yet, they may not realize exactly what they have to do on the test. Any test will be clearer to the students and easier for them to negotiate if it has clear headings and directions.

The *solutions* involve making sure the headings are distinct from the rest of the text (in the sense that they are italicized, made bold, or otherwise emphasized) and ensuring that they clearly indicate heading levels with different forms of placement and emphasis like those used in this article (left-justified title-caps and bold italics used for main headings and beginning of paragraph first letter cap with a period and bold italics for second-level headings).

In addition, given that the students taking these tests are usually second language speakers of the target language, the directions should probably be in the students' mother tongue, or if that is not possible, the directions should be simple and direct in the target language (with clear options for asking the teacher for further clarification). Two types of directions will often serve best: general and specific directions. General directions typically provide information to students about the overall test and apply to all sections of the test. Specific directions are particular to the section for which they are supplied. One thing to keep in mind: if the phrase or sentence appears in all of the specific directions, it probably belongs in the general directions.

Test Development Practices

In test development practices, teachers sometimes have problems with: proofreading the test, using a sufficient number of items, and examining student performances on the items.

Proofread the test. Another *problem* is that, even when a good deal of effort has gone into writing good quality items, clearly organizing those items, and providing clear headings and directions, other problems may still persist including typos, spelling errors, unclear formatting, and other problems that will make the test harder for the students to understand.

The *solution*, or at least a partial solution, is to carefully proofread the test several times even though you think you have finished it. I like to proofread my way through the test in different ways: reading from left to right on each line, then right to left; reading from top to bottom, and then bottom to top; I even throw the paper on the floor and look it over while standing above it (especially for logical formatting, e.g., making sure each item is on one page, that each reading passage is visible at the same time as the items associated with it; etc.). The trick is to look at the test from various perspectives because that will help in spotting typos and other problems before the tests are reproduced and handed out to students.

I also find that it helps to get others involved in the proofreading process because of the different and useful perspectives they may bring to the task. What I am suggesting is that you have a colleague, a former student, or even a spouse also proofread the test. You will be amazed at the sorts of problems they will uncover because their different perspectives on the test allow them to see things you are too close to the test to notice. Remember that, ultimately, when you administer the test to say 20 students, you will also have 20 people proofreading your test—people who are more than willing to point out a mistake that the teacher made in writing the test.

Use a sufficient number of items. The *problem* that some teachers create is that they try to test their course objectives with too few items. It stands to reason that more observations of a given phenomenon will be more accurate than fewer observations. This principle is well established in the sciences. However, even in language testing, common sense tells us that testing students with one multiple-choice item would not be reliable or accurate, indeed it simply wouldn't seem fair. Would two items be better? Or 3 items, or 10? So the principle that more items are generally better makes sense. The only real question is how many items are necessary to make the assessment of students reliable, accurate, and fair. The answer to that question depends on how good the items are. If the items are of good quality and suitable for the students in terms of their general proficiency level and what they are being taught, then fewer items will be necessary.

One *solution* is to make sure you have enough items to start with (say 50% more than you think you will need) so you can get rid of some items if they don't work very well. How many items should you have on your test? That will depend on common sense and thinking about the time constraints and the types of things you are asking your students to do on the items. So the end number will be different for each situation. But this I know, more items will generally do a better job of measuring what your students can do, but you can get away with fewer items if they are good ones.

Examine students' performances on the items. Another *problem* that arises for teachers is that they do not analyze their students' performance on their test items, much less revise those items. As a result, such teachers continue to use the same items or types of items over and over again even though those items do not work very well. You have probably found yourself in situations administering a test, when suddenly a student asks if there are two possible answers for number 11, and you realize she is right; then another student asks if any answer is correct for number 25, and you realize that there really isn't. So you tell the students to select the "best" answer, which essentially means that you recognize that there are problems with those items, and perhaps others. The next semester you are using the same test, when suddenly a student asks if there were two possible answers for number 11, and you instantly realize that you forgot to fix the items that had problems, even though students had helped you spot those problems.

One obvious *solution* is to carefully listen to students questions and comments about your test and take notes, then, after scoring the test, immediately take a few minutes to revise the test and save that version in such a way that you will remember to use it the next time you test the same material.

A more *systematic solution* would be to consider the first administration of any test a pilot run. You can then analyze the results statistically and revise on the basis of what you learn from the analysis. The actual item analyses that are probably most appropriate for classroom tests are called the *difference index*, which "shows the gain, or difference in performance, on each item between the pretest and posttest" (Brown, 2003, p. 18) and the *B index*, which "shows how well each item is contributing to the pass/fail decisions that are often made with CRTs" (p. 20). These item analysis statistics are both based on the simple percentage of students who answered each item correctly at different times or in different groups. Using these statistics and common sense, you can select those items that are most closely related to what your students are leaning in your course, replace any items that are not closely related, and make

fairer decisions based on your test scores. For more about the steps in calculating and interpreting these classroom-test item statistics, see Brown (2003, 2005). If you take the time to do item analysis every time you administer a test, your tests will continue to get better every time you use them.

Test Validation Practices

In test validation practices, teachers sometimes have problems with: reporting the scores as percentages, checking the reliability of the test, and thinking about the validity of the test.

Report the scores as percentages. The first validity-related *problem* is that some teachers report the number of items answered correctly to students along with information about the distribution of scores (e.g., the high and low scores, the number of students at each score, etc.). Teachers probably do this because they (and their students) are thinking in terms of the bell curve. This approach will lead students to think competitively in terms of how they did relative to other students, rather than to how much learning they were able to demonstrate on the test.

The *solution* is a simple one. In order to encourage the students to think about how much they have learned, report their scores as percentages and explain to them that the scores reveal what proportion they learned of the material taught in the course. Your score report will be even more informative if you can give students their percentage scores for each section of the test or for each objective in the course. The important thing to keep in mind for yourself and your students is that your classroom tests are designed to measure their learning in the course (criterion-referenced testing), not to spread them out on a continuum (which is norm-referenced testing like that done on standardized tests).

Check the reliability of the test scores. The *problem* here is that some teachers fail to think about or check the degree to which they might be making decisions about their students (grading, passing/failing, etc.) based on unreliable information. What does reliability mean when it comes to test scores? Reliability can be defined as the degree to which a set of scores are consistent. This concept is important because teachers generally want to be fair and make decisions for all students in the same way. If the scores on a test are not consistent across time, across items, or especially across students, then the decision making may not be the same each time for all students. Thus reliability is really a question of fairness.

One *solution* to this reliability issue is for teachers to think about reliability in terms of *sufficiency of information*: "What teachers really need to know, from a reliability perspective, is, 'Do I have enough information here to make a reasonable decision about this student with regard to this domain of information?' The essential reliability issue is: Is there enough information here?" (Smith, 2003, p. 30). While Smith was pondering the idea of creating a reliability index for such an interpretation, teachers might simply ask themselves one question: do I have enough good quality information from these test items to make responsible decisions about my students?

Another *solution* to this reliability issue would be to directly address the question: To what degree are the scores on my test reliable? This could be addressed by calculating a reliability coefficient. These coefficients typically range from .00 to 1.00, which can be interpreted as a range from zero reliability to 100% reliability. Thus if a coefficient for a set of scores turns out to be .80, that means that the scores are 80% reliable (and by extension 20% unreliable). So generally, the higher this value is the more reliable the scores are. Most reliability estimates were designed for standardized tests and are not appropriate for classroom testing, but one such estimate, the Kuder-Richardson formula 21 (known affectionately as K-R21) is appropriate for classroom testing (as explained in Brown, 2005, p. 209). Calculating this coefficient is relatively easy, requiring only that the teacher first calculate the mean (*M*),

standard deviation (*SD*), and number of items (*k*) (all of which can be calculated fairly easily in the Excel spreadsheet program), then enter these values into Walker's calculator for K-R21 at:

http://www.cedu.niu.edu/~walker/calculators/kr.asp

The result will be a reliability coefficient that the teacher can interpret as an indication of the consistency of the test scores involved.

Think about common sense validity issues. Another *problem* that some teachers have is that they fail to consider the validity of their test scores. Validity has traditionally been defined as the degree to which a set of test scores is measuring what it was intended to measure. In recent years, language testers have expanded their thinking about validity to include issues related to the consequences and values implications of how those scores are used.

Classroom teachers who wish to address validity issues need not get involved in learning elaborate theories or statistical procedures. They can instead start by asking themselves the following relatively simple questions (adapted from and explained more fully in Brown, 2012):

1. How much does the content of my test items match the objectives of the class & the material covered?
2. To what degree do my course objectives meet the needs of the students?
3. To what degree do my test scores show that my students are learning something in my course?
4. Will my students think my test items match the material I am teaching them?
5. How do the values that underlie my test scores match my values? My students' values? Their parents' values? My boss' values? Etc.?
6. What are the consequences of the decisions I base on my test scores for my students, their parents, me, my boss, etc.?

Your answers to the above questions will probably be matters of degree, but they will nonetheless help you understand the degree to which your test scores are valid.

Conclusion

In this column, I have explored some of the problems that teachers may face in their classroom testing in terms of test writing practices, test development practices, and test validation practices. These notions are elaborated in Table 1 which shows the three general categories of testing practices (writing, development, and validation) and the general suggestions made in this column, but also summarizes the solutions offered for ways to implement those suggestions.

If even a few teachers begin to use a few of these suggestions, I have no doubt that their testing and therefore their teaching will improve. As a result, they will be better serving their students, themselves, and their institutions.

Table 1. ***Summary of Practices, Problems, and Solutions in Classroom Testing***

Practices	Problems	Solutions
Test Writing	Some teachers allow too little time for writing their test items (perhaps because they lack training in item writing)	*Create good quality items* by getting ahold of a good book on language testing and reading up on what good quality items are and how to write them; be sure to allot sufficient time by starting early.
	Tests sometimes seem like a disorganized hodge-podge of items	*Organize the items* by keeping items that are testing the same language point together; grouping items of the same format (e.g., multiple-choice, true-false, etc.); and keeping reading or listening items together with their passages.
	Students may find the organization of a test confusing, or worse, they may not understand what they need to do	*Provide clear headings and directions* by emphasizing headings (using bold, italics, etc.) and using them hierarchically; writing directions in students' mother tongue or in very simple/clear English; and using general and specific directions.
Development	Even with all of the above, other problems may remain (e.g., typos, spelling errors, etc.)	*Proofread the test* carefully yourself and get others to do so as well (including perhaps a colleague, former student, or even spouse) because another set of eyes can spot things you are too close to the test to see.
	Some teachers try to test their course objectives with too few items	*Use a sufficient number of items* by always writing 50% more good quality items than you think you will need; use common sense in deciding how many items to use while taking into account time constraints and the nature of the items.
	Some teachers fail to analyze and revise items even though they will use them again	*Examine the students' performances on the items* by listening to their questions/ comments during the test and revising; by considering the first administration a pilot test and performing item analysis (i.e., the *difference index* and *B index*) and revising.
Validation	Some teachers report the number of items correct and explain scores in terms of the bell curve	*Report the scores as percentages* and explain to students that the scores reveal how much they learned of the material taught in the course, rather than how the scores spread them out.
	Some teachers fail to consider & check if their score-based decisions are founded on unreliable information	*Check the reliability of the test items* in terms of sufficiency of information (the degree to which you have enough information to make consistent decisions) and calculate and interpret a K-R21 reliability coefficient.
	Some teachers fail to consider & check the validity of the scores on their tests	*Think about common sense validity issues* in terms of the degree to which the scores are measuring what you intended and the consequences/implications of your score uses by answering the six validity questions posed above.

References

Brown, J. D. (2003). Questions and answers about language testing statistics: Criterion-referenced item analysis (The difference index and B-index). *SHIKEN: The JALT Testing & Evaluation SIG Newsletter, 7*(3), 18-24). Accessed online September 29, 2013 at http://jalt.org/test/bro_18.htm

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment* (New edition). New York: McGraw-Hill.

Brown, J. D. (2012). What teachers need to know about test analysis. In C. Coombe, S. J. Stoynoff, P. Davidson, & B. O'Sullivan (Eds.), *The Cambridge guide to language assessment* (pp. 105-112). Cambridge, Cambridge University.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University.

Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford: Oxford University.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and practice, 22*(4), 26-33.

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA

Your question can remain anonymous if you so desire.

Upcoming Language Testing Events

The 5th Association of Language Testers in Europe (ALTE) International Conference: April 10 – 14, 2014

Abstract submission deadline: (closed)

Venue: Maison Internationale, Cité Internationale Universitaire de Paris, Paris, France

Conference homepage: http://events.cambridgeenglish.org/alte-2014/index.html

The 11th European Association for Language Testing and Assessment (EALTA) Conference: May 29 – June 1, 2014

Abstract submission deadline: (closed)

Venue: University of Warwick, Coventry, UK

Conference homepage: http://www2.warwick.ac.uk/fac/soc/al/research/conferences/ealta2014

The 36th Language Testing Research Colloquium (LTRC): June 4 – 6, 2014

Abstract submission deadline: (closed)

Venue: VU University Amsterdam, The Netherlands

Conference homepage: http://ltrc2014.nl/

2014 Pacific Rim Objective Measurement Symposium (PROMS): August 2 – 6, 2014

Abstract submission deadline: April 26, 2014

Venue: Guangzhou, China

Conference homepage: http://www.confchina.com/index.html

The Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) Conference: November 27 – 29, 2014

Abstract submission deadline: (not yet open)

Venue: The University of Queensland, Brisbane, Australia

Conference homepage: http://www.altaanz.org/altaanz-conferences.html

Shiken Research Bulletin Editorial Board

Submissions

If you have a paper that you would like to publish in *Shiken Research Bulletin,* please email it in Microsoft Word format to the General Editor at:

jaltteval+srb@gmail.com