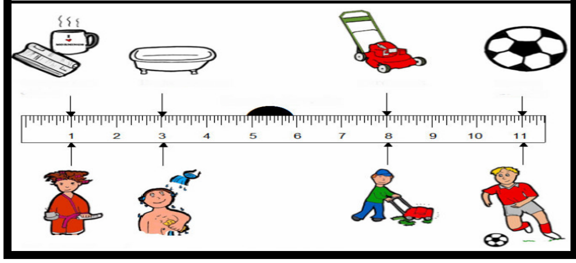
Meaningful Construct Maps



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 21 No. 2 Autumn 2007 ISSN 1051-0796

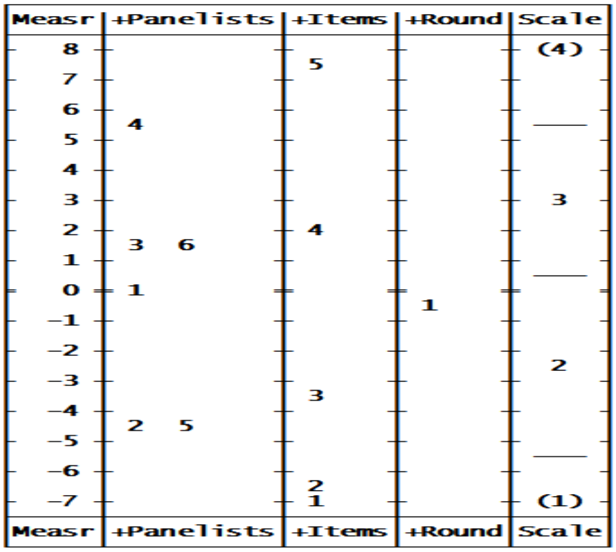
Evaluating Bookmark Judgments

There has been a great deal of work done on how to evaluate standard-setting procedures. Hambleton and Pitoniak (2006) suggested procedural, internal, and external criteria for evaluating standard-setting methods. Procedural criteria focus on implementation issues and documentation, internal criteria stress inter-panelist and intra-panelist consistency, and external criteria address comparisons to other methods and the reasonableness of the performance levels.

The two most popular methods for collecting judgments from standard-setting panelists are modified-Angoff and the bookmark procedure (Cizek and Bunch, 2007). The bookmark procedure (Mitzel, Lewis, Patz, and Green, 2001) is becoming the standard-setting method of choice in many statewide assessment programs, even though there has been less research conducted on bookmark methods as compared to modified-Angoff methods (Plake, 2007).

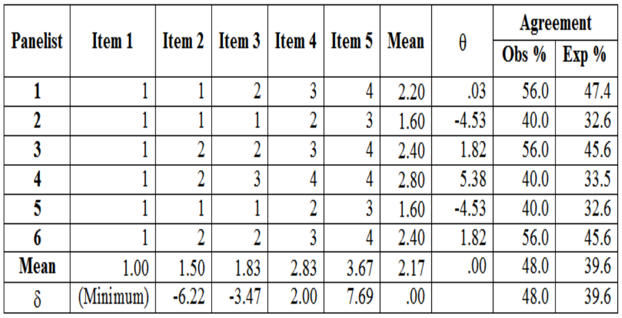
In a series of articles with my colleagues, I proposed using Rasch measurement theory to evaluate the quality of judgments obtained from standard-setting panelists (Engelhard & Anderson, 1998, Engelhard & Cramer 1997, Engelhard & Gordon, 2000, Engelhard & Stone, 1998). A summary of this approach is forthcoming (Engelhard, in press). This approach is based on the many-faceted Rasch (MFRM) model, and it incorporates many of the internal criteria described by Hambleton and Pitioniak (2006). The MFRM model can be used to evaluate the quality of standard-setting judgments obtained from bookmark panelists. The MFRM model for bookmark judgments is:

Figure 1. Variable Map of Ferdous-Plake (2007) Example



```
|Measr|+Panelists|+Items|+Round|Scale|
+  8  +          +      +      + (4) +
|     |          |  5   |      |     |
+  7  +          +      +      +     +
+  6  +          +      +      +     +
|     |  4       |      |      | --- |
+  5  +          +      +      +     +
+  4  +          +      +      +     +
+  3  +          +      +      +  3  +
+  2  +          +  4   +      +     +
|     |  3   6   |      |      |     |
+  1  +          +      +      +     +
|     |          |      |      | --- |
*  0  *  1       *      *      *     *
|     |          |      |  1   |     |
+ -1  +          +      +      +     +
+ -2  +          +      +      +     +
|     |          |      |      |  2  |
+ -3  +          +      +      +     +
|     |          |  3   |      |     |
+ -4  +          +      +      +     +
|     |  2   5   |      |      |     |
+ -5  +          +      +      +     +
|     |          |      |      | --- |
+ -6  +          +      +      +     +
|     |          |  2   |      |     |
+ -7  +          +  1   +      + (1) +
|Measr|+Panelists|+Items|+Round|Scale|
```

Table 1. Data matrix and summary of MFR analyses (One Round)



Panelist	Item 1	Item 2	Item 3	Item 4	Item 5	Mean	θ	Agreement	
								Obs %	Exp %
1	1	1	2	3	4	2.20	.03	56.0	47.4
2	1	1	1	2	3	1.60	-4.53	40.0	32.6
3	1	2	2	3	4	2.40	1.82	56.0	45.6
4	1	2	3	4	4	2.80	5.38	40.0	33.5
5	1	1	1	2	3	1.60	-4.53	40.0	32.6
6	1	2	2	3	4	2.40	1.82	56.0	45.6
Mean	1.00	1.50	1.83	2.83	3.67	2.17	.00	48.0	39.6
δ	(Minimum)	-6.22	-3.47	2.00	7.69	.00		48.0	39.6

Reliability of separation for items (R=.94) and for panelists (R=.81)

$$\text{Log } [P_{nijk} \;/\; P_{nij(k-1)}] = \theta_n - d_i - w_j - t_k \qquad [1]$$

where

P_{nijk} = probability of panelist n giving a bookmark rating of k on item i for round j,

$P_{nij(k-1)}$ = probability of panelist n giving a bookmark rating of k-1 on item i for round j,

θ_n = judged performance level for panelist n,

d_i = judged difficulty for item i,

w_j = judged performance level for round j, and

t_k = judged performance standard for bookmark rating category k relative to category k-1. The rating category coefficients, t_k, defines the performance standards or cut scores.

Table of Contents

In order to illustrate the MFR model, an example from Ferdous and Plake (2007) is presented in Table 1. There are six panelists providing bookmark ratings (performance levels from 1 to 4) for five items. The cell entries represent panelist judgments regarding the performance level of each item. The observed means for the items range from 1.00 to 3.67 reflecting the ordered items that would be listed in the ordered item booklet. The observed judgments range from 1.60 to 2.80 with Panelists 2 and 5 having the lowest view of performance and Panelist 4 with the most severe judgments of performance needed to succeed on these five items. This ordering is reflected in the estimated values for the θ's and the d's.

This information is presented in the variable map in Figure 1. Both panelists and items are centered at zero, and round (only one round in the example) is not centered. The panelists range in interjudge agreement from 40.0% to 56.0%. The overall observed agreement is 48.0% with an expected agreement of 39.6% based on the model. Item 1 is not included in the agreement statistics because all of the panelists agreed to rate it in category 1.

Table 2. Category Statistics

Category	Usage	Performance Standards (Cut scores)	
		Measure	SEM
1	21%	-	-
2	33%	-5.82	1.03
3	25%	.28	1.03
4	21%	5.53	1.22

Table 2 presents the category statistics. Within the framework described here, the measures for the category coefficients are defined as the performance standards or cut scores. This definition provides the opportunity to use several graphical displays for practitioners to understand panelist judgments. Figure 2 shows the category probability curves.

Ferdous and Plake (2007) report an interjudge inconsistency index of 36%. If we report this as a consistency or agreement index, then the value is 64%. This value is higher than the Rasch estimate of 48.0% because Item 1 is included in their estimates of interjudge consistency. The MFR model provides the opportunity to go beyond a single index of inter-judge consistency. It also makes available an array of model-data fit indices and graphical displays for exploring more deeply judgments of panelists using the bookmark procedure.

Additional work is currently underway to explore the utility of this approach for evaluating bookmark ratings in a variety of standard-setting situations. Experience is still needed to determine whether or not the MFR model can provide a suite of internal criteria for examining bookmark judgments obtained from standard-setting panelists.

George Engelhard, Jr.
Emory University

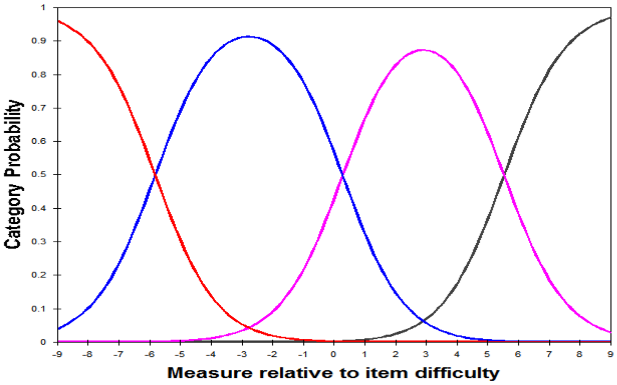


Figure 2. Category Probability Curves

Cizek, G.J., & Bunch, M.B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage

Engelhard, G. (in press). Evaluating the judgments of standard-setting panelists using Rasch measurement theory. In E. V. Smith, Jr., and G. E. Stone (Eds.), Applications of Rasch measurement in criterion-referenced testing, JAM Press.

Engelhard, G., & Anderson, D.W. (1998). A binomial trials model for examining the ratings of standard-setting judges. Applied Measurement in Educ., 11(3), 209-230.

Engelhard, G., & Cramer, S. (1997). Using Rasch Measurement to evaluate the ratings of standard-setting judges. In M. Wilson, G. Engelhard, & K. Draney. (Eds.). Objective Measurement: Theory into Practice, Volume 4 (pp. 97-112). Norwood, NJ: Ablex.

Engelhard, G., & Gordon, B. (2000). Setting and evaluating performance standards for high stakes writing assessments. In M. Wilson & G. Engelhard (Eds.), Objective Measurement: Theory into Practice, Volume 5 (pp. 3-14). Stamford, CT: Ablex.

Engelhard, G., & Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. Educ. and Psychological Measurement, 58(2), 179-196.

Ferdous, A., & Plake, B. (2007). Interjudge inconsistency index for body of work, yes/no, and bookmark standard setting procedures. Retrieved September 2, 2007: http://www.unl.edu/buros/biaco/pdf/pres07ferdous01.pdf

Hambleton, R C., & Pitoniak, M.J. Setting performance standards. In R. Brennan (Ed.), Educational Measurement, 4th Ed. (pp. 433-470) Westport, CT: Praeger Publishers.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed), Setting performance standards: Concepts, methods and perspectives (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Assoc.

Plake, B.S. (2007, April). Standard setters: Stand up and take a stand! 2006 career award address presented at the annual NCME meeting, Chicago, IL.

IOMW 2008
March 22-23, 2008 - New York

The final date for paper and symposium applications is January 18, 2008. The final program will be announced on February 1, 2008. Further information on IOMW 2008 and on the paper and symposium submission process can be found at

http://www.jampress.org/

Scroll down the JAM home page until you arrive at IOMW 2008 (near the bottom) and click on that link. This will take you to the IOMW 2008 page. Once on this page, please click on the appropriate links to download a printable PDF form for paper or symposium submission.

Data Recognition Corporation and JAM Press are pleased to announce that IOMW 2008 will be held in New York City at New York University on March 22 and 23, 2008, just prior to the AERA 2008 annual meeting. The meeting will be held in the Helen and Martin Kimmel Center for University Life located at 60 Washington Square South in New York City. This location is a short bus, cab, or train ride from the AERA conference hotels.

The two-day program will allow for a total of 80 presentations in 16 sessions. There will be two plenary sessions and 6 concurrent sessions each day with a maximum of 5 presenters in each session. Each session will be 90 minutes in length, allowing each presenter approximately 14 minutes per presentation, with 20 minutes at the end for questions and answers. Individual papers will be grouped into a session with papers sharing common themes. Symposia based on 4 to 5 commonly themed papers will be considered.

This is the fourteenth meeting of IOMW, a series of biannual meetings that originated in 1981. The first IOMW was organized by Ben Wright and held at the University of Chicago. We all hope that this will be a special IOMW, located very near Ben Wright's childhood home in Greenwich Village. We are organizing a session to talk about Ben's New York experiences.

A conference dinner is also planned.

Richard M. Smith, Editor
Journal of Applied Measurement

The Difficulty of an MCQ Item

"We shall define the difficulty of a multiple choice test item as being a function of that proportion of individuals answering the item which knows which of the alternatives is the best answer. This definition involves the assumption that there is some objective criterion which determines that one particular alternative is a better answer to the item than any of the others."

Paul Horst, "The Difficulty of a Multiple Choice Test Item" Journal of Educational Psychology, xxiv (1933), 229-232.

Rasch Measurement SIG
Ballot Announcement and Call for Officer (Self-)Nominations

At the AERA Annual Meeting, March 24-28 in New York City, a new SIG Secretary/Treasurer and SIG Chair will take office. If you are interested in one of these positions or would like to nominate someone for one of these positions, please contact Ed Wolfe, *edwolfe~at~vt.edu*

It is time to **elect new officers for the Rasch SIG**. We need to conduct an election by email ballot by December 24th, 2007, and we need to elect a SIG Chair and Secretary/Treasurer. The term is for two years, and any current member of the Rasch SIG is eligible for election as an officer. The current officers have served a single term, so both are eligible for re-election. To avoid low voter turnout due to the end of the academic semester the election will be held by email during the first week of December. If you would like to nominate someone or self-nominate for a SIG office, please email Ed Wolfe (edwolfe~at~vt.edu) identifying that individual (with email address) and the office for which you would like to nominate the individual. At that time, he will confirm that the individual desires to serve and ask that those agreeing to serve provide him with a brief biographical description **no later than November 26th, 2007.** Ed will then assemble the ballot and distribute it via email.

Another issue that will appear on that ballot is a proposal for a decrease in the **annual Rasch SIG dues**. Dues for the Rasch SIG are currently $15 for one year or $25 for two years-a rate that is slightly higher than average for AERA SIGs (typical rates are $5 to $10 per year, although some are higher than ours). The SIG dues rate was set when the *Rasch Measurement Transactions* were being mailed to each member, and the SIG needed to cover the cost of printing and mailing of that newsletter. Currently, our balance is about $7,700. Our annual expenditures include the following (approximate values):Website ($200), Annual Meeting rentals ($350), AERA fees ($225)-about $800 per year. Our current membership is about 180 members, which is up only slightly from April of 2007. At that rate, we have an income of less than $2700 in dues each year - "less than" because a small percentage chooses to pay for two years of dues at a time. Clearly, we do not currently have a need for the current dues levels. Even if we were to reduce dues to $5 per year, we would have sufficient income (about $900 per year) to cover our current spending needs.

The current officers (Tom O'Neill and Ed Wolfe) would like to add a voting item to the officer election ballot to reduce the amount of the annual dues. Their current thinking is to reduce the amount to $10 per year and $20 per two years, but they'd like your thoughts on that proposal before calling the vote. Please email thoughts on this issue to Ed Wolfe, edwolfe~at~vt.edu

Understanding Lexiles

Often when trying to discuss the development of reading proficiency, measurement specialists and reading specialists seem to be talking at cross-purposes. There may be more to the issue than either perspective recognizes. Reverting to argument by metaphor, measurement specialists are talking about measuring weight; reading specialists, about providing proper nutrition.

There is a great deal involved in physical development that is not captured when we measure a child's weight and the process of measuring weight tells us nothing about whether the result is good, bad, normal, try to schedule a doctor's appointment, or go to the emergency room without changing your clothes. Evaluation of the result is an analysis that comes after the measurement and depends on the result being a measure. No one would suggest that, because it doesn't define nutrition, weight is not worth measuring or that it is politically sensitive to talk about in front of nutritionists. A high number does not imply good nutrition nor does a low number imply poor nutrition. However, a measurement of weight is always a part of any assessment of well-being.

A Lexile score, applied to people, is a measure of reading ability, which is taken to mean the capability to make meaning from words and sentences. Lexiles, as applied to text, is a measure of how difficult it is to make meaning from that text. A colleague of mine offered as a counter example Hemingway's "For Whom the Bell tolls" (840L). Since a 50 percentile sixth grade reader could self engage with this book, something must be wrong because the book was written for adults. This counter-example of an instance where Lexiles "do not work", if true, is an interesting case. I have two counter-arguments: one, all measuring instruments have limitations to their use and, two, Lexiles may actually be describing Hemingway appropriately.

First, outside the context of Lexiles, there is always difficulty in scoring exceptional, highly creative writing for both humans and computer algorithms. (I would venture to guess that many publishers, who make their livings recognizing good writing, would reject Hemingway, Joyce, or Faulkner-like manuscripts if they received them from unknown authors.) I don't think it follows that we should avoid trying to evaluate exceptional writing. But we do need to know the limits of our instruments.

I rely, on a daily basis, on a bathroom scale. I rely on it even though I believe I shouldn't use it on the moon, under water, or for elephants. It does not undermine the validity of Lexiles in general to discover an extraordinary case for which it does not apply, if that is in fact the case. Again, we need to know the limits of our instrument.

Second, given that we have defined the Lexile for a text as the difficulty of decoding the words and sentences, the Lexile analyzer may be doing exactly what it should with a Hemingway text. Decoding the words and sentences in Hemingway is not that hard: the vocabulary is relatively simple, the sentences relatively short. The Lexile score will reflect that.

Understanding and appreciating Hemingway is something else again. I am trying to make a distinction between reading ability and reading comprehension. You have to be able to read before you can comprehend what you have read. Analogously, you have to be able to do arithmetic before you can solve math word problems. The latter requires the former but the former does not guarantee the latter.

The Lexile metric is a true developmental scale that is not related to instructional methods or materials, or to grade-level content standards. The metric reflects increasing ability to read, in the narrow sense, increasingly complex text. As students advance through our reading/language arts curriculum, they should progress up the Lexile scale. Effective standards-based instruction should cause them to progress on the Lexile scale; analogously good nutrition should cause children to progress on the weight scale.

One could coach children to progress on the weight scale in ways counter to good nutrition. One might subvert Lexile measurements by coaching students to write like Hemingway, on one end, or like Supreme Court Justice Antonin Scalia, on the other. There need to be other checks to ensure that we are effecting what we set out to effect. This does not invalidate either weight or reading ability as useful things to measure.

There are many things in the curriculum that are not assessed directly by the Lexile analyzer. Understanding imagery and literary devices, locating topic sentences and main ideas, recognizing sarcasm or satire, comparing authors' purposes in two passages would not be considered in the Lexile measure. The role of standards-based assessment is to identify which constituents of reading ability and reading comprehension are present or absent.

The role of the Lexile measure is to provide a measure of the student's status on a narrowly defined, interval scale that extends over the length of reading from Dick and Jane to Scalia and Roberts. It does not define reading, recognize the breadth of the ELA curriculum, or replace grade-level content standards-based assessment, but it can help us understand the results of the assessment and help us design instruction appropriate to the student. On the one hand, we cannot expect students to say anything intelligent about text they cannot decode, nor should we attempt to assess their analytic skills using that text. On the other hand, we should expect to assess and improve their analytical skills using text they can decode.

Ronald Mead
Data Recognition Corp.

HKSoQOL 2008

***May 15-18,* Guangzhou, China.**

The Hong Kong Society for Quality of Life is co-organizing the 2008 Asian Chinese Quality of Life Conference with the First Affiliated Hospital of the Guangzhou University of Traditional Chinese Medicine and the School of Public Health of the Sun Yat-Sen University. There will be plenary, symposium, free paper and workshop sessions and will cover topics including (1) Health related QOL research findings and methodology; (2) QOL in rehabilitation and social services; (3) QOL research in Chinese Medicine; (4) QOL research in Nursing care.

http://www.hksoqol.org/conf2008

We have invited more than 30 overseas and local QOL researchers to lecture in the plenary and symposium sessions. **Prof. Trevor Bond will deliver a 3-hour workshop on RASCH and address in a symposium session.** This will be a valuable opportunity for you to update your knowledge on QOL research. This year, we have special symposium and free paper sessions for QOL research in psychiatry, palliative care, rehabilitation, nursing, Chinese medicine and social services.

The conference will also be a good platform for you to share your QOL research findings and experiences with fellow the workers. You may submit an abstract for free paper or poster presentation on-line at the conference web site. You may also apply for scholarship if your abstract is accepted.

Please note that you can enjoy a **great reduction of registration fee if you register before end of 2007.** Please visit the conference web site for details.

http://www.hksoqol.org/conf2008

Looking forward to sharing with you in the conference.

Kwok Fai Leung
Chairman
Hong Kong Society for Quality of Life
conference~at~hksoqol.org

Effect of Misfit on Measures

Question: I have a fairly large sample of 5,000 subjects. As an experiment I ran the calibration with all subjects and then again with the 500 worst fitting (OUTFIT mean-square range from 2 to 9.9) subjects excluded. There was some change in parameter estimates and item fit, but not huge, not what I expected. This is comforting, but has this been the experience of others or is it probably a quirk of my data or the large sample size?

Answer: Yes, your experience with trimming misfitting persons is typical. You are removing the most unpredictable, the noisiest part of the data, so the remaining data must have a slightly more orderly, closer-to-Guttman pattern. So expect to see a slight increase in the logit range of the measure estimates when you trim the data. But it is unusual for this slightly wider spread of the measures to have any substantive implications except where subject measures are adjacent to pre-set cut-points.

Third International Rasch Measurement Conference

Perth, Western Australia

22-24 January 2008

Pre-Sessions: Jan. 7-11, 14-18, 21, 2008

Topics for the conference:

- Cumulative models for attitude and trait measurement-dichotomous and ordered category models.
- Unfolding models for preference and choice -folding the Rasch models
- Rasch model applications in education (e.g., large scale test equating, benchmarking), psychology (e.g., intelligence testing, linking quantitative and stage developmental data)
- Item banking
- Computer adaptive testing
- Marketing (e.g., pairwise designs for preference and choice studies)
- Health care outcomes (e.g., linking performance scales)
- Using simulation studies for clarifying methodological issues (e.g., tests of fit, measurement precision, effects of multidimensionality and response dependence)
- Developments in Rasch modeling (e.g. differential item functioning)
- Understanding response processes compatible with the Rasch models
- Epistemology, fundamental measurement and Rasch models
- History and philosophy of measurement and Rasch models

January 7-11 Introductory course on Rasch measurement. Includes use of the program RUMM

January 12 Course barbecue

January 14-18 Advanced course in Rasch measurement. Includes use of the programs RUMM, RATEFOLD

January 21 One day workshop focusing on using RUMM

January 22-24 Conference papers on applications of Rasch and related measurement models in any substantive field of application - education, psychology, health care and rehabilitation, marketing, etc.

January 22 *Conference dinner* at the Nedlands Golf Club, located two miles from the city of Perth, and overlooking the Swan River.

http://www.education.uwa.edu.au/httpwww.education.uwa.edu.aunews/rasch_conference

An Example of Grader Consistency using the Multi-Facet Model

The issue of consistent grader severity is an on-going concern for all who score performance examinations. This study explored the consistency of common grader severity across three performance examination administrations. Each performance examination administration was analyzed using the multi-facet Rasch model which produced calibrations of grader severity.

The data are from three annual administrations of a medical oral examination labeled administrations A, B, and C. Between administrations, there were some common graders and some non-common graders. To be included in the study, a common grader had to rate candidates in at least two of the three administrations, although some graders were common to all three administrations. In this study, there were 115 common graders who met this criterion. This examination also had standardized items and tasks which graders used to rate the candidates. The candidates for each of the three administrations were completely different; however, the examination process was the same.

Graders rate a random sample of the candidates who take the examination in a given administration. During the course of each examination administration each grader gives many ratings which are used to calibrate his/her severity. Because so many ratings are given by each examiner, the calibrations of grader leniency or severity are very precise.

The items in this oral examination were carefully developed for consistency and content coverage. The skills being rated were well defined and the same across all administrations. The rating scale is well defined for each rating level. Graders were trained prior to the examination with regard to the content of the items and examination procedures. Many of the graders have a great deal of experience in the examination process. The multi-facet formula used for this analysis was:

$$\log (P_{nijkx} / P_{nijk(x-1)}) = B_n - D_i - C_j - H_k - F_x$$

where B_n = ability of candidate *n;*
D_i = difficulty of item *i;*
C_j = severity of grader j;
H_k = difficulty of task *k;* and
F_x = Rasch-Andrich threshold or step calibration.

Because the examination materials are so well standardized, differences in grader severity within examination administrations are most likely due to inherent differences in grader expectations and standards, which will probably not change substantially due to training. Grader severity was calibrated using the multi-facets model for each of the three examination administrations. The center of each scale was anchored at 0.00 logits for all three exam administrations. Next the grader severity calibrations were compared across examination administrations using z-scores and correlations for the common graders.

Using the grader severity estimates and their measurement errors, the standardized difference between grader severities across administrations was calculated using z-scores (Forsyth., Sarsangjan, and Gilmer, 1981). The formula used to obtain standardized differences for grader severity calibrations is:

$$Z_j = (C_{j1}-C_{j2})/(S_{j1}^2+S_{j2}^2)^{½}$$

where C_{j1} and C_{j2} are grader severity estimates for each administration, and S_{j1} and S_{j2} are the estimated measurement errors associated with these severity estimates.

Correlations were also used to confirm the patterns of grader severity.

The calibrated severity estimates for the common graders ranged from -1.78 to1.55 logits during administration A, from -2.07 to 1.50 logits during administration B and from -1.96 to 1.52 logits during administration C. Within each examination administration, the severity estimates among graders were significantly different from each other as indicated by a Chi-Square test and a Separation reliability. This difference in grader severity was significant even after training and working within a carefully structured examination process.

An absolute z-score of 1.96 or greater, indicates 95% confidence that there is a statistically significant difference in grader severity across administrations. Comparison of the grader severity estimates across administrations using the z-score analysis found that of the 115 common graders, only one was statistically significantly different in severity across administrations at the 95% confidence level. The common grader who was significantly different was very lenient during administration A, but significantly more severe during administrations B and C.

The graders within an administration were significantly different from each other in severity; however, they were consistent within themselves within and across examination administrations. This suggests that severity is a grader characteristic that should be included in the analysis of performance examinations to improve validity and reliability. The multi-facet model provides the opportunity to incorporate this facet into analysis of performance examinations and to better understand grader grading patterns.

Mary E. Lunz
Measurement Research Associates, Inc.
http://www.measurementresearch.com/

Forsyth., Sarsangjan, and Gilmer, 1981, Forsyth, R., Sarsangjan, V. and Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. Applied Psychological Measurement, 5, 175-186.

Global Rasch Fit Statistic

Question: A Journal Editor insists I include a global statistic for fit of the Rasch model to my data. What do you recommend?

Answer: The Editor misunderstands the Rasch model, but this is not the moment to rectify that. Numerous global fit tests have been proposed reflecting the different ways in which the data can misfit the unattainable ideal of the Rasch model. Here’s a practical approach. For each observation, there is a standardized residual and a model probability. So we can always compute usefully approximate chi-square statistics, regardless of missing data:

1. Pearson chi-square = sum of squared standardized residuals for all observations.

2. Log-likelihood chi-square = -2 * sum of the natural logarithms of the model probabilities for all observations.

In practice, these values will differ. So we can choose the value better fitting our intentions, as is usually done in statistical modeling, or report both statistics. In both cases, the degrees of freedom for dichotomous data approximate:

d.f. = data point count - (person count + item count)

Omit items and persons with zero or perfect scores before doing these computations. For polytomies, also deduct from the d.f. the number of active categories (less 2) for each polytomous scale.

Since the expectation of a chi-square statistic is its d.f., you can obtain a more accurate estimate of the d.f. by simulating multiple sets of data with the same measurement structure as your data, and then using the average of their chi-square values as the reported d.f. for your chi-square.

Noise and Random Error

Question: In Rasch analysis, how does noise differ from random error?

Answer: Every observation is conceptualized to consist of three components:

1. Its expected value. This is the amount predicted from the Rasch model and the parameter estimates (ability, difficulty and rating scale structure).

2. Model randomness or modeled random error. This is the randomness in the data predicted by the Rasch model, which is a probabilistic model. It is the Bernoulli binomial variance or multinomial variance, “the model variance of the observation around its expectation”. The Rasch model uses this for estimating the distance between the parameter estimates, the Rasch measures.

3. Unmodeled randomness. This is the part of each observation that contradicts the Rasch model. It makes the mean-square statistics depart from 1.0. We don't want this randomness because it degrades measurement. From the perspective of the Rasch model, this component is random, i.e., unpredictable, but it may be highly predictable from other perspectives, e.g., “Robin has a response set.”

Statistically, “noise” is “2.+3.”, but often we use “noise” to mean “3.” or even “2.”. If there is obvious ambiguity, we use terms like “modeled randomness” for “2.”, and “unmodeled noise” for “3.”.

There is the paradoxical situation that some of the “3. Unmodeled randomness” can cancel out some of the “2. Model randomness” This happens when the data overfit the model, and the mean-squares are less than 1.0. So sometimes, “noise” only refers to the part of “3. Unmodeled randomness” that adds to the model randomness in the observations.

Rasch-related Coming Events

Nov. 26-29, 2007, Mon.-Thurs. AARE Conference (W. Fisher), Australia, http://www.aare.edu.au/

Dec. 2007 - Dec 2008 3-day Rasch courses, (A. Tennant, RUMM) , Leeds, UK, http://home.btconnect.com/Psylab_at_Leeds/Courses.htm

Jan. 7-11, 2008, Mon.-Fri. Introductory course on Rasch measurement, (Andrich, RUMM), Australia, http://www.education.uwa.edu.au/httpwww.education.uwa.edu.aunews/rasch_conference

Jan. 14-18, 2008, Mon.-Fri. Advanced course on Rasch measurement, (Andrich, RUMM), Australia, http://www.education.uwa.edu.au/httpwww.education.uwa.edu.aunews/rasch_conference

Jan. 21, 2008, Mon. One-day RUMM Workshop, (Andrich, RUMM), Australia, http://www.education.uwa.edu.au/httpwww.education.uwa.edu.aunews/rasch_conference

Jan. 22-24, 2008, Tues.-Thurs. 3rd International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch models, Australia, http://www.education.uwa.edu.au/httpwww.education.uwa.edu.aunews/rasch_conference

Feb. 15 - March 15, 2008, Fri.-Fri. Practical Rasch Measurement with Winsteps online course, (M. Linacre), www.statistics.com/courses/rasch

March 22-23, 2008, Sat.-Sun. IOMW 2008, New York, http://www.jampress.org/

March 24-28, 2008, Mon.-Fri. AERA Annual Meeting, New York, www.aera.net

May 2-30, 2008, Fri.-Fri. Many-Facet Rasch Measurement online course, (M. Linacre, Facets), http://www.statistics.com/courses/facets

May 15-18, 2008, Thur.-Sun. 2008 Asian Chinese Quality of Life Conference (T. Bond), China http://www.hksoqol.org/conf2008

July 28 - Nov. 22, 2008 Introduction to Rasch Measurement and Traditional Test Theory online course (D. Andrich, RUMM2020), http://www.education.uwa.edu.au

What Use Are Measures?

Question: I've spent a lot of time and effort estimating Rasch measures. Now what do I do with them?

Reply: Rasch measures can be used wherever person raw scores and percentages or item *p-values* would be used in conventional test reporting and statistical analysis. Rasch measures have the linear properties that most statistical routines (and non-specialist readers) assume of your numbers, but which raw scores and p-values don't have. So you can use Rasch measures for reports, plots, descriptive statistics, statistical tests, regressions, etc.

A powerful use of Rasch measures is to draw pictures (item and person maps) which show the item hierarchy (the construct validity) of the items, and the person hierarchy (the predictive validity) of the persons.

The hierarchy of item difficulties is especially important because it defines what is being measured, the measurement ruler. Does the ordering of the items in difficulty match the intentions of the instrument developer and the expectations of those planning to use the test results? It is yet more instructive if, prior to data collection, the test development team sketch out the intended difficulty order of the items. This can then be compared with the order estimated from the data. The comparison usually confirms most of the intended ordering, so supporting the validity of the test. But the comparison may also point out an item or two that were supposed to be easy but are not, and vice-versa. This leads to a better understanding of the underlying construct, the latent variable, and also to improvement in the items. When an a supposedly easy item is reported to be difficult in practice, this can also indicate an area where better education or training is needed of those for whom the test or assessment is intended.

For instance, in the Knox Cube Test (a standard dataset, Wright & Stone, 1979), a gap in the item hierarchy indicates where new items should be written to target the sample. The item hierarchy also indicates how "number of taps", "number of reversals", and "length of sequences" affect item difficulty, so leading to a better understanding of how we store information in our short-term memory.

Item difficulty measures are important if the instrument is to be used for setting criterion-level cut-points. They are also crucial for equating instruments, and for selecting items for adaptive administration.

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
www.rasch.org/rmt
Editor: John Michael Linacre
Copyright © 2007 Rasch Measurement SIG
Permission to copy is granted.
SIG Chair: Thomas O'Neill, Secretary: Ed Wolfe
Program Chairs: Sharon Solloway & Ed Wolfe
SIG website: http://www.raschsig.org/

Varying Item Discrimination = Multidimensionality?

Question: Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31, 19-26,* states that "test scaling models are self-contradictory if they assert both unidimensionality and different slopes for the item characteristic curves." Do differences in item discrimination always indicate multidimensionality?

Answer: In situations like this, it is helpful to think of parallels in physical measurement. Suppose we are measuring length with old-fashioned cloth tape-measures. These can become stretched along parts of their range. If we compared measurements of lengths with two of these stretched tape measures, we would see that, to start with, they would say the same numbers. Then the less-stretched tape measure would have higher numbers, i.e., be more discriminating. Then they might agree again. Then the other tape measure might have higher numbers. Length is unidimensional, but the "tape measure ICCs" cross, perhaps several times along their lengths. We could call "stretching", i.e., changes of length-discrimination, another dimension, in the same sense as "guessing" is another dimension. But these are not usually what is mean by "multidimensionality".

On the other hand, we might have two good cloth tape measures, but they might not always be parallel or straight. They might "snake" somewhat as we use them. Again they would sometimes agree and sometimes disagree due to crisscrossing "tape measure ICCs". Here we could agree that the problem is "multidimensionality". The tape measures are not in a straight line.

Fit Statistics: Size or Significance?

Question: Which one is most relevant to decide if an item is misfitting, the size of the mean-square statistic or its statistical significance?

Answer: When considering measurement dilemmas, it is always helpful to think of the equivalent situation in physical measurement. The statistical significance reports how certain we are that the measurement misrepresents with the data - but not how serious the misrepresentation is. The mean-square reports the size of the misrepresentation, but not how certain we are that this isn't merely reflecting the random component in the data predicted by the Rasch model.

In physical measurement, we are usually more concerned about the size of any possible misrepresentation ("measure twice, cut once") than about how certain we are that there is a misrepresentation ("I'm sure I measured it right, so there's no need to measure it again!"). If size of misrepresentation is more important than certainty, then the size of the mean-square is more crucial than its significance. But much of statistics is based on hypothesis testing, where only the probability of misrepresentation is seriously considered.