Rating Score Data Analysis by Classical Test Theory and Many-Facet Rasch Model

Tsai-Wei Huang
National Chiayi University,
Chiayi, Taiwan

Gwo-Jen Guo
National Changhua
University of Education,
Changhua, Taiwan

William Loadman
The Ohio State University,
Columbus, USA

Fang-Mei Law
Tiffin University, Tiffin,
USA

More and more judgments conducted by raters are involved in psychological and educational assessment and testing situations. Two measurement theories, CTT (classical test theory) and MFRM (many-facet Rasch model), were implemented in the analyses of rating scale data. Parameters of item difficulty, item discrimination, and reliability were compared between CTT and MFRM. One hundred and twenty-four applying proposals were evaluated by 64 raters on a 24-item rating form. The results showed that item difficulty indices estimated by CTT and by MFRM were highly consistent, but item discrimination indices were remarkably uncorrelated. The values of reliability provided by CTT and MFRM were very consistent.

Keywords: rating score, classical test theory, many-facet Rasch model

Introduction

Recently, more and more judgments conducted by raters are involved in psychological and educational assessment and testing situations (Han, Kreiter, Park, & Ferguso, 2006; Angkaw, Tran, & Haaga, 2006; Wilder, Braaten, Wilhite, & Algozzine, 2006; Holmberg, de Geer, Swärd, & Fernell, 2004; Kulinna & Zhu, 2001). In some educational situations, as teachers need to evaluate students' performance through some tasks, the scores would involve at least three facets, the rater (teachers), the examinee (students), and the task (test). In this case, raters need to judge the optimum of answers provided by examinees, and scores earned by examinees are based on the judgments of raters. Obviously, it is not appropriate to assess the students' performance only through the examinee and item facets and ignore the influence of raters' leniency/severity on examinees' scores as we interpret in multiple-choice tests. Therefore, some item characteristics such as item difficulty, item discrimination, and even test reliability are highly correlated with raters' judgments. However, many conventional analyses discard the effect of the raters and simply focus on the interaction between examinees and tasks. The factor of rater indeed needs to be taken into consideration in item analysis. Thus, the main purpose of this study intends to compare item characteristics on rating scores between CTT (classical test theory) and MFRM (many-facet Rasch model).

CTT

CTT is well known more than 150 years. It has contributed to three remarkable measurement

---

Tsai-Wei Huang, professor, Department of Counseling, National Chiayi University.

Gwo-Jen Guo, professor, Department of Guidance and Counseling, National Changhua University of Education.

William Loadman, Faculty Emeritus, Quantitative Research, Evaluation and Measurement, The Ohio State University.

Fang-Mei Law, associate professor, Psychology and Criminal Justice, Tiffin University.

achievements: “A recognition of the presence of errors in measurements, a conception of error as a random variable, and a conception of correlation and how to index it” (Traub, 1997, p. 8). In fact, CTT is a measurement theory about test scores that introduces three concepts—test score, true score, and error score (Hambleton & Jones, 1993). True score and error score are theoretical concepts and unobserved. The test score is the only score that can be observed. The relationship between these three scores is assumed as linear, i.e., test score is equal to true score plus error score. Alternatively, true score can be expressed as the expected scores in parallel tests.

CTT assumes that no correlations exist between true score and random error, neither between two random errors nor between true score and random error in different tests. An advantage of CTT is that they are based on relatively weak assumptions, i.e., they are easy to meet in real test data. However, CTT encounters a problem of interactive dependence between item and person factors. An examinee with a fixed ability level would be scored low on difficult tests and would be scored high on easy tests (Hambleton & Jones, 1993). Conversely, an item with fixed difficulty level would be regarded as difficult if most examinees fail on it and as easy if most examinees succeed on it. This limitation is because CTT cannot separate an ability parameter from the difficulty parameter.

Another source of dissatisfaction with CTT lies on the assumption of equal standard errors of measurement across all examinees. Nevertheless, since scores on any test are unequally precise measures for examinees of different ability, as Hambleton, Swaminathan, and Rogers (1991) indicated, the assumption of equal errors of measurement for all examinees is implausible.

Traditional item analyses focus on the analysis of item difficulty, item discrimination, and item distraction. In rating scores, since no answer is absolutely right or wrong and scores are based on a rater’s judgment, it is not appropriate to calculate item distraction. However, these item analysis approaches can be revised so as to calculate item difficulty and item discrimination in rating scored items. A high value of item difficulty indicates that the particular item tends to be rated by high scores, and vice versa. The high values of an item discrimination index, reporting the difference between the proportions of high and low scorers, are reputed to flag good items, and low values bad ones (Lopez & Stone, 1998).

Noll, Scannell, and Craig (1979) provided two indices for item difficulty (P) and item discrimination (D) in a rating-score situation. Both of them rely on the ratios of the scores in different groups and the maximal and minimal scores in an item. These two revised formulas are defined as follows:

$$P = \frac{S_H + S_L - (2N \times X_{MIN})}{2N(X_{MAX} - X_{MIN})} \quad (1)$$

and

$$D = \frac{S_H - S_L}{N(X_{MAX} - X_{MIN})} \quad (2)$$

where N represents of 25% total number persons, X_{MAX} and X_{MIN} are the highest and lowest scores in an item, respectively, and finally, S_H and S_L indicate the sum of scores in the high- and low-score groups, respectively.

On the other hand, the reliability of test scores on rating scales provided in CTT reflects the level of

internal consistency. Cronbach α represents the general case of internal-consistency reliability. It is equal to the complement of the proportion of individual item variances to total score variance, and was adjusted by the number of items (Allen & Yen, 1979).

MFRM

MFRM is an extension of the Rasch model. It is especially useful in the situations in which more than two facets interact to produce an observation (Linacre, 1994). As an extension of the Rasch model, MFRM shares the same assumptions with the Rasch family. First, the assumption of unidimensionality indicates that only one dimension of variable is measured in the process of assessment. Without holding this assumption well, as Linacre (1994) stated that many misfits between the data and model will occur if the assumption is rigidly applied. The second assumption is the local independence. It implies that, after taking judges' severity into account, no relationship exists between judges' ratings for different examinees on different items. These two assumptions are similar to those proposed by the general item response models. In addition, MFRM assumes that "judges are modeled to differ in severity or leniency... (and that) judges are modeled to exhibit some degree of stochastic behavior when awarding ratings" (pp. 6-8). The former statement indicates that severity is a parameter characterizing the judge, just as the ability characterizing an examinee or the difficulty characterizing an item. The latter statement illustrates that the estimation of the severity parameter could contain measurement errors, which could lead to misfits to the model.

Based on these assumptions, MFRM inherits two major features of Rasch model. First, the estimates of parameters are invariant across facets. In a three-facet Rasch model, for instance, the ability parameter for the examinee facet, the leniency/severity parameter for the judge facet, and the ease/difficulty parameter for the item facet are independent of each others. They keep their invariance property when they interact with other parameters. Second, these facet-free characteristics permit the facets to be linearly added. Thus, the facets can be expressed on a common scale and share the same property on the linear scale (Linacre, 1994).

MFRM not only provides an analysis of item difficulty, but also provides analyses of person ability and rater severity. The parameter of item difficulty is separated from that of examinee ability and from that of rater severity. Moreover, all of the measures provided in MFRM are based on new scale scores (logits), rather than on raw scores used in CTT.

Although the Rasch model further assumes that all items to have the same or approximately equal discriminations, Wright (1995) suggested that the infit *MnSq* (mean square statistics) can be regarded as an index providing information on item discrimination. The infit mean square statistic is an information-weighted Chi-square statistic divided by its degrees of freedom and is sensitive to unexpected inlying patterns among informative, on-target observations (Linacre & Wright, 1994). The expected value of the infit mean square is 1. Any value substantially less than 1 indicates that the observed item characteristic curve is flatter than the Rasch model and fails to differentiate different abilities levels of examinees. However, if the infit value is substantially larger than 1, it indicates an unusual discrimination and that must be investigated further (Wright & Stone, 1979).

On the other hand, for the MFRM and Rasch family, reliability is called separation reliability and is defined as the ratio of the squared separation to the sum of the squared separation and 1. Mathematically, the expression of separation reliability is equivalent to KR-20, or Cronbach Alpha (Linacre, 1999). The separation reliability shows how different the measures are. Notice that MFRM not only provides separation reliability for

the subjects, but also provides rater separation reliability and item separation reliability. The general form of reliability defined in the Rasch family is as $REL = SEP^2 / (1 + SEP^2)$, where SEP is the separation of parameter and is defined as the ratio of adjusted standard deviation to the root mean square standard error for the estimated parameter.

Method

Participants and Instrumentation

The data for this study were obtained from the scaled responses from the scoring of 124 competitive grant applications. There were 64 raters and each proposal was independently rated by a trio of raters using a 24-item scoring rubric. Raters were randomly assigned to each proposal and then rotated after each proposal rating so that the same three raters never rated an additional proposal together.

The raters were professionals in the content area of the proposals and were specifically recruited to participate in the scoring task. The raters were trained on proposal reading, the specifics of the RFP (request for proposal) and the scoring rubric prior to reading and scoring of the proposals. They were also pre-screened to avoid potential conflicts of interest in the proposals they were asked to score.

The 24-item scoring was created by referencing and indexing the major sections of the RFP, thus building in content validity. The items were scaled using a six-point continuum with a score of 1 representing “very weak”, “lowest rating” and/or “missing” and a score of 6 represented “exceptionally strong”, “exemplary”, or the “highest rating”. Each rater read and scored approximately 6 proposals. Thus, there were a total of 372 “reads” (124 times 3) and approximately 8,928 responses (372 times 24).

Analysis Design

Item difficulty and item discrimination. In CTT, the minimum score in the data set is 1 and 6 for the maximum score across all items. The number of trials (N) in this case can be regarded as the number of “reads” and, thus with a quarter selection, there were 93 high-score “reads” and 93 low-score “reads” (25%) chosen from the total “reads” (372). SPSS software was used to calculate the item difficulty index P and the item discrimination index D , which were provided by Noll et al. (1979).

In MFRM, based on the application, item and rater facets, a three-facet Rasch model were formulated in this study:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k \quad (3)$$

where P_{nijk} and P_{nijk-1} represents the probability of application n being awarded on item i by rater j a rating of k and $k-1$, respectively. B_n means the ability of the application n , D_i means the difficulty of item i , C_j means the severity of judge j , and F_k means the difficulty of the step up from category $k-1$ to category k . Parameters of item difficulty and item discrimination will be estimated through the FACETS, a Rasch measurement computer program (Linacre, 1999), under scales of logit and $MnSq$.

Reliability. In this study, the coefficient of Cronbach α was provided in CTT analysis. Three kinds of separation reliability (person, item, and rater) were provided in the analysis of MFRM.

CTT does not lend itself well to this situation... we usually have all raters scoring one common application or an estimate of the three raters for each application averaged across all trios of raters. This will have to be addressed.

Results

Descriptive Statistics

Descriptive statistics about item characteristics were displayed in Table 1. As can be seen, the means of rating scores range from 2.64 (AN17) to 4.75 (AN3), with AN20 showing the widest dispersion ($SD = 1.49$). Most items tend to be slightly negatively skewed, except for AN16, 17 and 19, which implies that applications tend to receive relatively higher ratings from raters in these negatively-skewed items, especially for item AN10 ($SK = -0.98$). Similarly, most items tend to be platykurtically distributed, but this is not apparent (the highest value of kurtosis is occurred in AN21, $KT = -0.95$).

Table 1

Summary of Statistics for Each Item ($N = 124 \times 3 = 372$)

Item #	M	SD	SK ^a	KT ^b	Item #	M	SD	SK ^a	KT ^b
AN1	4.15	1.22	-0.34	-0.53	AN13	3.92	1.31	-.47	-0.36
AN2	4.32	1.12	-0.63	0.19	AN14	4.32	1.16	-.83	0.38
AN3	4.75	1.01	-0.70	0.39	AN15	3.93	1.20	-.33	-0.31
AN4	4.46	1.05	-0.57	0.10	AN16	2.99	1.30	.14	-0.72
AN5	4.56	1.11	-0.62	-0.10	AN17	2.64	1.27	.44	-0.49
AN6	4.50	1.19	-0.65	-0.13	AN18	3.37	1.39	-.14	-0.75
AN7	4.39	1.07	-0.68	0.16	AN19	2.79	1.32	.34	-0.64
AN8	4.50	1.29	-0.90	0.29	AN20	3.54	1.49	-.18	-0.94
AN9	4.42	1.43	-0.82	-0.17	AN21	3.37	1.47	-.06	-0.95
AN10	4.64	1.22	-0.98	0.57	AN22	4.04	1.27	-.55	-0.33
AN11	4.33	1.14	-0.51	-0.22	AN23	4.08	1.29	-.51	-0.52
AN12	3.70	1.32	-0.22	-0.73	AN24	4.11	1.02	-.69	0.25

Notes. ^a = skewness; ^b = kurtosis.

To grasp the whole information of the three parameters (ability, item ease/difficulty, and rater leniency/severity), a map of calibrations for all facets was shown in Figure 1. As can be seen, there showed a negatively skewed distribution for application ability, but a slightly negatively skewed distribution for rater leniency/severity, and a slightly positively skewed distribution for item ease/difficulty. In general, all three distributions approximate a normal distribution. Paralleling with the expected scores in scale 1 (abbreviated as S1), most applications tended to gain high scores (from 3 to 5) on these relatively easy items although more than half of raters were calibrated on the severe side of the distribution. Note that this slightly positive distribution for item difficulty in the logit scale revealed the same information and meaning as that of the slightly negative skewness in the scale of the CTT in Table 1. It is also interesting that the range of application ability dispersed so widely (from +1.4 to -1.7 logits) that the most and least able applications were out of the ranges of rater severity (from +0.9 to -1.2 logits) and item difficulty (from +1.1 to -0.7 logits). This means that except for a few extreme applications, items are suitable for most applications who encountered consistent raters in this study.

Item Difficulty and Item Discrimination

The analyses of item difficulty are consistent across CTT and MFRM. The easiest item (AN3) identified by CTT with $p = 0.74$ is analogously identified as the easiest one by MFRM with item difficulty equal to -0.73 logits. Similarly, the item of AN17 is identified as the hardest one among all items both by CTT ($p = 0.33$) and by MFRM (item difficulty = 1.14 logits). Moreover, all items show consistent levels of difficulty analyzed by CTT and MFRM with a spearman rank correlation equal to 0.99 ($p < 0.01$). This consistency can also be demonstrated by a significant correlation (Pearson product-moment correlation, $r = -0.99$, $p < 0.01$).

Measr	Application	Rate	Item	S.1
+ 2 +				+(6) +
	512			
	405 409			5
	477 478 516			
	454		17	
+ 1 +	408 412 413 418 476 490 497		+ 19 +	+ +
	411 420 431 435 442 458 510	3032	16	
	410 425 451 466 482 489 494 502 515	3061		
	422 461 462 469 480 481 496 498 499 501 507 517 524			—
	429 434 438 444 459 473 479 485	3037 3047 3057		
	426 430 449 464 467 504	3011 3016 3020 3059	18	
	401 406 414 421 432 437 447 450 453 484 492 509	3003 3008 3042	21	
	445 457 472 483 511 518	3007 3039 3051 3052 3060	12 20	4
	428 486 493 508	3000 3015 3030 3031 3041 3044 3050 3056		
	404 443 452 463 468 470 500 506 520	3013 3028 3034 3036 3038 3040 3043	13 15	
* 0 *	465	3001 3035 3054 3055 3065	* 22 *	* *
	419 427 441 455 471 474 488 514	3006 3019 3025 3062	23 24	—
	403 415 417 440 448 487 495 519	3009 3014 3023 3058 3063 3066	1 11	
	407 416 436 513 522	3012 3021 3022 3024 3026 3045 3064	2 7 14	
		3004 3010 3033	4 5 6 9	
	433 460 491 503 505 521	3027	8 10	3
	423 424	3018 3029		
	456 475	3005	3	
	446	3049		
	402			—
+ -1 +				+ +
		3002		
		3017		
				2
	439			
	523			
+ -2 +				+(1) +
Measr	Application	Rate	Item	S.1

Figure 1. Map of ability, severity, and difficulty parameters.

Item Difficulty and Item Discrimination

The analyses of item difficulty are consistent across CTT and MFRM. The easiest item (AN3) identified by CTT with $P = .74$ is analogously identified as the easiest one by MFRM with item difficulty equal to -0.73 logits. Similarly, the item of AN17 is identified as the hardest one among all items both by CTT ($P = .33$) and by MFRM (item difficulty = 1.14 logits). Moreover, all items show consistent levels of difficulty analyzed by CTT and MFRM with a spearman rank correlation equal to .99 ($p < .01$). This consistency can also be demonstrated by a significant correlation (Pearson product-moment correlation, $r = -.99, p < .01$).

To more visualize this consistency, Figure 2 provided a scatter plot with a scale of p and logit provided by

CTT and MFRM, respectively. As can be seen, all items are almost in a straight line with opposite scale directions: the more the P value an item has, the less the logit value, and vice versa. This also indicated the consistency of item difficulty estimated by CTT and MFRM.

Figure 2. Scatter plot of scales for item difficulty from CTT(P) vs. MFRM ($logit$).

Figure 3. Scatter plot of scales for item discrimination from CTT (D) vs. MFRM ($MnSq$).

On the comparison of item discrimination, there were many inconsistencies between items in the two scales. Items with the poorest two discrimination indices in CTT estimation (AN3 and AN4 with $D = .27$ and $.28$, respectively) were ranked as the 11th and 7th poorest discrimination items in the MFRM analysis. On the other hand, items with the highest discrimination indices in the CTT estimation (all AN15, AN18, and AN21 with $D = .48$) were estimated as 0.7, 0.9, and 1.1 $MnSq$ but ranked divergently (see the 3rd, 10th, to 20th items). In addition, there was an extreme item (AN24) with the value of infit mean square ($MnSq = 0.5$,

substantially less than 1), recognized as not as able as the other items to differentiate applications' abilities, but was identified as a highly discriminating item in the analysis of CTT ($D = .43$). In contrast, an item identified as a fair discriminator (AN9) in the analysis of CTT ($D = .37$) was labeled as a noisy discriminator with a high infit value ($MnSq = 1.6$) by MFRM. A Pearson product-moment correlation ($r = -.082, p = .08$) and a Spearman rank correlation ($r_s = .17, p = .43$) demonstrated this inconsistency between the two discrimination estimations that showed a rough circle and a very asymmetric pattern in Figure 3.

Reliability

Table 2 reveals the values of estimated reliability associated with the analyses of CTT and MFRM. As can be seen, MFRM provides the highest level of reliability overwhelming the other method with item separation reliability ($REL. = .99$), application separation reliability ($REL. = .96$), and rater separation reliability ($REL. = .95$). The high reliability of the facet measures indicates that a strong difference of parameters exists among their corresponding facets. On the other hand, the Cronbach alpha coefficient provided by CTT is quite high ($\alpha = .93$). This implies that individual items are highly consistent with the total score.

Table 2

Comparisons of Reliability among CTT, G Theory, and MFRM

Analysis	CTT	MFRM
Reliability	.93 (Cronbach α)	.96 (Application) .95 (Rater) .99 (Item)

Discussion

Item difficulty indices estimated by CTT and by MFRM are highly consistent and correlated. This implies that both methods of estimating item difficulty index are equivalent. One of them can be used to interpret item difficulty for each item without missing much information. However, MFRM not only provides the estimate of item difficulty, it also provides the estimates of examinee ability and rater severity, which are separated from item difficulty. This consistent finding of difficulty estimates from CTT and MFRM is just as that Lord (1980) had mentioned. It is necessary to note that a two-parametric model used in Lord's book, but here is one parametric model.

In the comparison of item discrimination, many inconsistencies between CTT and MFRM may be due to two effects. First, CTT only considers the top quarter and the bottom quarter of samples. This implies that item discrimination in CTT is only appropriate for classifying the top quarter and the bottom quarter of examinees, but the item discrimination index cannot classify the middle part of examinees. However, MFRM analysis uses the entire data set to identify infit values. In this case, MFRM considers the middle part of the sample and provides corresponding infit indices. Second, CTT ignores the effect of measure error variations. However, MFRM considers the modeled variance in the estimation of infit statistics for each individual on each item. If the sum of individual variances is too large, the raw data set will depart from the ideal model for this item and, thus, may result in a low infit value. Conversely, if the sum of individual variances is too small, the raw data set will still depart from the Rasch model and might lead to a noisy infit value. Therefore, when we consider the total sample of examinees and individual modeled variance on each item, an item with a high discrimination index identified in the analysis of CTT might become more ambiguous in distinguishing examinees' abilities.

In the comparison of reliability, even though the mathematical formations of reliability introduced by the two methods are equivalent (the proportion of the variance of random error to the variance of total), their meanings are different. The Cronbach alpha coefficient emphasizes that the internal consistency between individual items and total score. A high coefficient value signifies means that a high level of consistency. High-total-score earners tend to earn high scores on individual items, and vice versa. However, it only provides the consistency information for items, with variability of item ease/difficulty and rater leniency/severity, neither for examinees nor for rater embedded in the error variance.

The high values of MFRM separation reliability shown in this study reveal two facts. First, the variance of random error is substantially smaller than the variance of applications, raters, and even items and, thus, most of the variety among applications, raters, or items is due to the main effects of their differences, rather than the effect of random error; Second, the high reliability of application measures indicates that the values of ability are strongly different across applications. Similarly, the high reliability of rater measures indicates that there is a strong distinction in rater leniency/severity among the raters, and the approximately perfect reliability of item measures indicates that there is a very apparent differentiation of item difficulty among these items. By not taking the rater leniency/severity into account in the analysis, some unadjusted application scores will be unjustly advantaged (lenient raters) or disadvantaged (severe raters).

To sum, we found similar estimates of item difficulty, reliability but different estimates of item discrimination by CTT and MFRM. MFRM provides more refined and defensible information. CTT has the variability associated with rater leniency/severity and variability of the applications embedded into the error term. That is, the separation of the estimates of parameters from MFRM seems to provide more diagnostic information for examinees, items, and even raters than the CTT theory does. In addition, it is more defensible than the CTT method because it takes the non-independence of the raters into consideration in the analysis, which cannot be easily handled by the CTT method. In fact, CTT only provides information for items. However, in the data set, some raters appear in application and not necessary to appear in other applications. Thus, it sounds a good suggestion that MFRM provides a better way to get a rating-base performance assessment than CTT does.

References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, C.A.: Brooks/Cole Publishing Company.
- Angkaw, A. C., Tran, G. Q., & Haaga, D. A. (2006). Effects of training intensity on observers' ratings of anxiety, social skills, and alcohol-specific coping skills. *Behaviour Research & Therapy*, 44, 533-544.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, C.A.: Sage Publications.
- Han, J. J., Kreiter, C. D., Park, H., & Ferguso, K. J. (2006). An experimental comparison of rater performance on an SP-based clinical skills exam. *Teaching & Learning in Medicine*, 18, 304-309.
- Holmberg, K., de Geer, L., Swärd, C., & Fernell, E. (2004). Behavioural and learning problems in schoolchildren related to cognitive test data. *Acta Paediatrica*, 93, 976-981.
- Linacre, J. M. (1994). *Many-Facet Rasch measurement* (2nd ed.). Chicago, I.L.: MESA.
- Linacre, J. M. (1999). *A user's guide to FACETS: Rasch measurement computer program*. Chicago, I.L.: MESA.
- Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, 8(2), 360.
- Lopez, W. A., & Stone, M. H. (1998). Raters & rating scale. *Popular Measurement*, 1(1), 60-65.

- Lord, F. M. (1980). *Applications of item response theory to practice testing problem*. Hillsdale, N.J.: Erlbaum Publishers.
- Kulinna, P. H., & Zhu, W. (2001). Fitness portfolio calibration for first-through six-grade children. *Research Quarterly for Exercise and Sport*, 72, 324-334.
- Noll, V. H., Scannell, D. P., & Craig, R. C. (1979). *Introduction to educational measurement* (4th ed.). Boston: Houghton Mifflin.
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4), 8-14.
- Wilder, L. K., Braaten, S., Wilhite, K., & Algozzine, B. (2006). Concurrent validity of the strength-based “behavioral objective sequence”. *Journal of Applied School Psychology*, 22, 125-139.
- Wright, B. D. (1995). 3PL or Rasch. *Rasch Measurement Transactions*, 9(1), 408-409.
- Wright, B. D., & Stone, M. H., (1979). *Best test design: Rasch measurement*. Chicago, I.L.:MESA.