CLEAR Exam Review

Volume XIX, Number 1
Spring 2008

A Journal

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 403 Marquis Ave., Suite 200, Lexington, KY 40502.

Editing and composition of this journal have been written by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact Stephanie Thompson at (859) 269-1802, or at her e-mail address, sthompson@mis.net, for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting Janet Horne at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

Coeditor
Michael Rosenfeld, Ph.D.
Educational Testing Service
Princeton, NJ 08541-0001
mrosenfeld@ets.org

Coeditor
F. Jay Breyer, Ph.D.
Prometric
2000 Lenox Drive
Lawrenceville, NJ 08648
jay.breyer@prometric.com

CLEAR Exam Review

VOLUME XIX, NUMBER 1 SPRING 2008

Contents

 ISSN 1076-8025

Investigation of the Item Characteristics of Innovative Item Formats

ANNE WENDT, PHD, RN, CAE
National Council of State Boards of Nursing, Inc.

Acknowledgements
Jerry Gorham, PhD, Psychometrician at Pearson VUE assisted with this research.
Kathy Gialluca, PhD, Psychometrician at Pearson VUE has reviewed this document.

Abstract

Advances in computer-based testing and Item Response Theory have created opportunities for the National Nursing Licensure Examination For Registered Nurses (NCLEX-RN®) to explore innovative items. This article compares traditional multiple-choice items with some innovative formats such as fill-in-the-blank calculation items, fill-in-the-blank ordered response items and multiple response items. Using two experimental datasets that were created from two time periods when the innovative items were pretested, items were calibrated using the Rasch (1PL) measurement model. Results of this study indicate that innovative items offer measurement properties that are comparable to or at times better than traditional multiple-choice items.

Introduction

Over a decade ago (1994) the U.S. National Nursing Licensure Examinations (NCLEX-RN®) moved from paper-and-pencil format using standard, four-option multiple-choice questions (MCQs) to Computerized Adaptive Technology (CAT) using that same item format. At that time, it was postulated that computers have the potential to assess new skills and abilities that have been difficult or extremely expensive to measure via traditional testing formats (McHenry & Schmitt, 1994). Innovations in computer-based testing include item types with features that include sound, graphics, animation and video integrated into the item stem, response options or both. In addition, use of Item Response Theory (IRT) has allowed the creation of measurement scales that are independent of the particular sample of people or test items used to create the scales (Lord & Novick, 1968; Lord, 1980). Furthermore, the use of IRT has facilitated the introduction of CAT for testing programs. With the introduction of CAT and innovative items, one research issue that is important to address is whether the innovative item types behave in ways that are comparable to the current MCQ item types. Comparability in item characteristics is important for issues such as model-data fit, scaling, and dimensionality.[1,2] Comparability in terms of how much time it takes an examinee to respond to an item (item response times), how many pretest items meet NCLEX statistical criteria (item survival rates), and other characteristics will be important for item production and test administration policy issues.

[1]The Rasch model assumes that there is one underlying dimension that is being measured such as nursing ability. Items must fit the model.
[2]Model-data fit refers to how well data (items) fit the Rasch measurement model. There are statistical indices generated by the Rasch model which help diagnose model-data fit (or "misfit"). For more information the reader should reference Best Test Design by Wright and Stone.

When introducing new item formats, one major concern is with dimensionality. A second major concern is with model-data fit. These are somewhat related issues, but it is possible for items to measure one major factor (dimension) and yet fit the models slightly differently. Large systematic problems with model-data fit ("misfit") might indicate some problem with dimensionality. This paper will focus on issues of model-data fit rather than issues of dimensionality because the design for this study, using current pretest data, did not produce data that would permit a reliable look at dimensionality. Thus, the purpose of this paper is to examine whether the innovative items are similar to MCQs in terms of their item statistical characteristics.

Methodology

To approach the issue of model-data fit, large datasets that are available from ongoing pretesting of the alternate (innovative) item types are used in this study. Two experimental datasets were used for NCLEX-RN from two testing periods in 2005. Each dataset included a combination of MCQ and innovative item types.

The NCLEX-RN examination is a variable length CAT examination. Each registered nurse (RN) examinee receives 15 pretest items in a CAT examination that may range from a total of 75 to 265 items in length. Items are calibrated using a Rasch (1 parameter logistic (1PL)) model (Rasch, 1980; Wright & Stone, 1979).

The first step in examining the innovative item types used in this study is to compare innovative item characteristics with known characteristics of MCQs. A question that needs to be answered is, are there perceptible differences in the statistical characteristics of items based on item type? Two approaches were used to examine the item characteristics: classical item analyses and calibration with the Rasch measurement model. Although the principal model used for calibration of items on the NCLEX examinations is the Rasch model, classical item analysis is also used in item screening procedures to eliminate items that do not show sufficient discrimination.[3]

The following are the types of innovative items under investigation in this study:

Fill-in-the-blank (FB)

Fill-in-the-blank items are examples of constructed response items where, unlike the selected response of MCQs, the examinee is not given a list of responses from which to choose the correct answer. An example of this type of item is the 'calculation item.' Nursing proficiency in calculation is a vital aspect of medication administration including calculation of medication doses and parenteral administration. In addition, nurses need to know how to calculate the client's intake and output. As can be seen from Figure 1, Fill-in-the-blank Calculation Item (FBC), the examinee is required to perform a calculation and then type the correct answer into the box/space provided.

> The nurse is monitoring the dietary intake and output of a client. The nurse observes that the client has consumed 8 ounces of apple juice, one hamburger on a bun, one-half cup of green beans, 8 ounces of tea, and one cup of ice cream. How many milliliters should the nurse record for the client's intake?
>
> _____milliliters

FIGURE 1. Fill-in-the-blank Calculation Item

Another fill-in-the-blank item type used within this examination is the ordered response sequence item, which is labeled (FBS).[4] In FBS items the examinee is required to sequence or rank order the options. For example, examinees are presented with a list of essential steps to a nursing procedure (e.g. cardiopulmonary resuscitation-CPR) and asked to rank order them in the correct sequence in accordance with established rules and guidelines. After deciding upon the correct sequence, the examinee lists the numbers in the correct order in the answer box or space provided. Figure 2, Fill-in-the-blank Ordered Response (FBS), illustrates this type of item.

> The nurse is caring for a client with an acute exacerbation of Crohn's disease. In what order would the nurse perform an abdominal assessment? Prioritize the nursing actions by typing the number of the **first** action the nurse should take, followed by subsequent actions in the answer space provided.
>
1	Test for rebound tenderness
2	Percussion
3	Auscultation
4	Palpation
5	Inspection

FIGURE 2. Fill-in-the-blank Ordered Response Item

[3]The calibration of an item is the difficulty of the item indicated by B-value using the Rasch model and a P-value (proportion of examinees answering the item correctly) using Classical statistics.
[4] FBS items for the NCLEX examinations have been revised and re-pretested as drag-and-drop items.

Multiple Response (MR)
While traditional MCQs allow the examinee to select a response from a list of four options, the multiple response innovative item is a variant on this item type that allows the examinee to choose one or more of the options provided (e.g., options 1,3, and 6). Figure 3 is an example of a Multiple Response item. This item format is used without cueing the examinee to the actual number of correct responses. Additionally, this format requires that the examinee have the ability to discriminate from a list of important content which has implications for examinees ability to think critically (Jodoin, 2003). Within nursing content, this item type is intended to identify the examinee's ability to consider all possibilities in providing client care in a given situation. Depending on the phrasing of the content in the item, an examinee may be required to discriminate between non-mutually exclusive actions that would impact the outcome of client care.

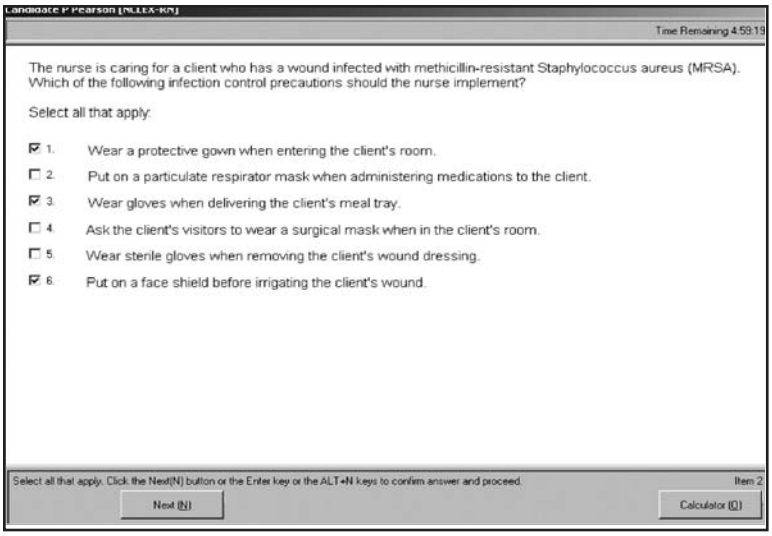

FIGURE 3. Multiple Response Item

According to Haladyna, multiple response items are similar to the MCQs used on the NCLEX examinations. Haladyna (1984) has stated:

> Strictly speaking, this format is the conventional multiple-choice where more than one right answer exists. The underlying rationale for this is that different lines of reasoning by test takers may logically lead to the selection of other answers that are also correct. The only material difference [is that] test takers are informed that they may select more than one answer but if they choose incorrectly a penalty is assessed (p. 47).

Parshall, et al, believe that the goal of selected response item formats is to improve measurement in some sense where innovative formats may tap slightly different cognitive constructs than do MCQs. For example, the ordering and multiple response item types may add a level of complexity to the task of responding (Parshall et al, 2000). Research done by Bennett, Morley, Quardt, Rock, Singley, Katz, & Nhouyvanisvong (1999) evaluated a computer-delivered response type for measuring quantitative skill, noting:

> Results showed that 'generating examples' scores were reasonably reliable but only moderately related to the GRE quantitative section, suggesting the two tests might be tapping somewhat different skills. Item features that increased difficulty included asking examinees to supply more than one correct answer and to identify whether an item was solvable. (p. 233)

This study examined the characteristics of these types of innovative items as compared to MC items.

Results

Classical Item Analysis
Table 1 provides a summary of item statistics by item type (format). The four item formats are MCQ (multiple choice, 4-option items), FBC (Fill-in-the-Blank Calculation items), FBS (Fill-in-the-Blank Sequenced/Ordered Response items, MR (Multiple Response items, or items that require an examinee to select all the answer options that apply). Note that the category "MC Anchors" is also shown for items that are re-pretested but are used primarily for the purposes of scaling the new pretest items to a common mean and standard deviation. Also note that in these data, only a few FBS items were pretested.

The mean sample sizes for the item calibrations met or exceeded standard calibration sample size targets and ranged from 489 to 560 for these RN items. As can be seen by Table 1, the MC Anchors show slightly higher

TABLE 1. Summary Statistics by Item Type for Two Datasets

		# items	Mean sample size	Mean PTBis	Mean Pvalue	Mean Response time (secs)	Mean Item Difficulty	Pct. Near Cut Score	# Items Failing Pretest	Pct. Items Failing Pretest
RN Items, DS 1	MC Anchors	97	495	0.11	0.60	59.8	-0.47	50.5%	28	28.9%
	FBC	92	491	0.14	0.76	174.8	-1.41	12.0%	12	13.0%
	FBS	8	489	0.07	0.43	116.6	0.37	12.5%	3	37.5%
	MC	824	490	0.08	0.68	54.5	-1.04	30.8%	394	47.8%
	MR	88	489	0.08	0.24	71.2	1.50	18.2%	40	45.5%
RN Items, DS 2	MC Anchors	15	560	0.11	0.58	63.7	-0.38	46.7%	2	13.3%
	FBC	12	549	0.21	0.67	170.4	-0.72	41.7%	1	8.3%
	MC	101	556	0.10	0.66	59.0	-0.98	35.6%	32	31.7%
	MR	16	549	0.10	0.24	78.1	1.43	18.8%	6	37.5%

point biserial correlation coefficients than the regular MCQ pretest items since these MC Anchor items are also operational items that have passed pretest based on previous statistical data. The mean point biserial correlation coefficients for the innovative item types FBC and MR are generally equal to or greater than the standard MCQ items. For the FBS items (only 8 items), the mean point biserial correlation coefficients are slightly lower than the standard MCQ items.

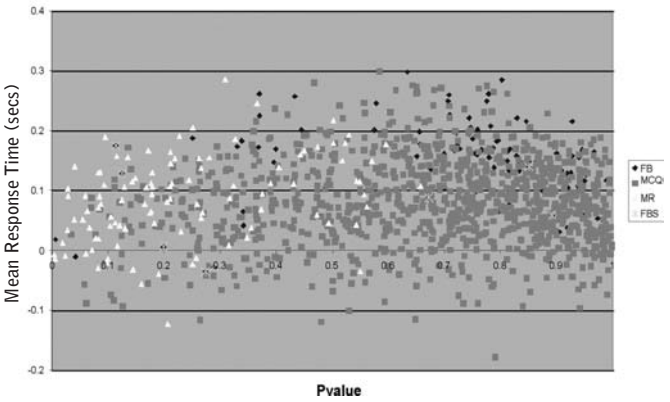

FIGURE 4. RN Items: Pvalue by Pt Biserial

Summary information for the mean item difficulties can be seen from the mean proportion correct column (P-value) or from the mean item difficulty column (Rasch B-value) in Table 1.[5] For the RN examination, the FBC items vary in difficulty compared to the MCQ items, while the FBS and MR items are consistently more difficult.

Mean item response time indicates that each of the innovative item types require more time on average for examinees to respond. While the standard MCQ items have traditionally required about a minute on average for the examinee to respond, the MR items require approximately 1 to 1.5 minutes, the FBS items require almost 2 minutes on average, and the FBC items require 2.5 to 3 minutes on average. However, it should be noted that the average response time for the MCQ calculation items for the RN examination is about 3 minutes on average, which is similar to the innovative FBC items.

Based on the last two columns of Table 1, the number and percentage of items failing the pretest screening, it appears that the survival rate for the new item types is at least as good, and may be better, than the survival rate for the standard MCQ item types. Results for the FBC item types are especially encouraging in their overall survival rates.

Figure 4 shows the relationship between the classical item statistics for the RN datasets (datasets 1 and 2 are combined). The figures plot the P-value by the point biserial for each of the RN items by item type. Note that the FBC item types have consistently higher point biserials than the MCQ items across the entire range of item difficulty. The MR items tend to have point biserials comparable to the MCQ items but tend to cluster toward the higher (more difficult) end of the scale.

[5]For the NCLEX examinations, the Rasch B-value provides an equal interval level of measurement and is an index of item difficulty that is centered at zero and generally ranges from -3.00 logits (very easy) to 3.00 logits (very difficult). For more information about the Rasch 1 PL model the reader is referred to Best Test Design referenced.

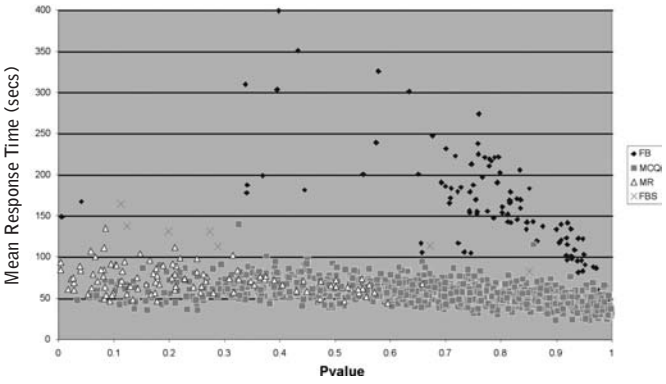

FIGURE 5. RN Items: Pvalue by Mean Response Time

Figure 5 shows the relationship between classical item difficulty (P-Value) and mean item response time for the RN items by item type. As might be expected, the MCQ items are fairly consistent in the amount of response time required per item, although there is a slight relationship between item difficulty and longer response time. The MR and FBS items show slightly higher overall item response times than the MCQ items, although these items tend to cluster toward the more difficult end of the scale. The FBC item types show a very strong relationship between item difficulty and item response time. That is, the more difficult the item, the more time required to complete the item.

1PL (Rasch) Analyses

Table 1 also shows means of the Rasch item difficulties by item type and dataset. As mentioned previously, the MR items seem to be very difficult, whereas the FBC items for this RN dataset appear to be somewhat easier than standard MCQ items. However, when the FBC items are compared to MCQ calculation items there seems to be no significant differences in difficulty.

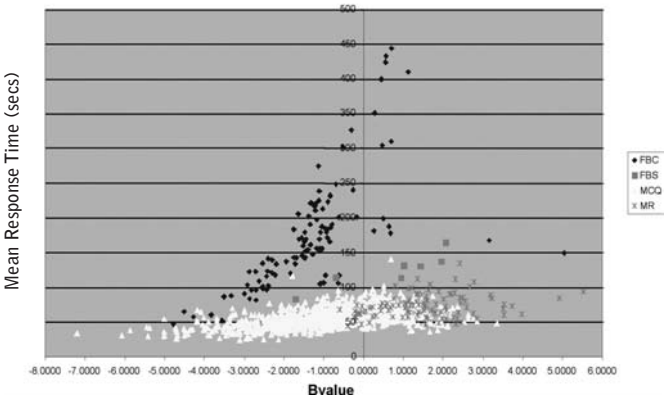

FIGURE 6. RN Items: Items Difficulty by Mean Response

Figure 6 shows the relationship between item difficulty and item response time even more clearly since item difficulty is plotted in terms of the Rasch b-parameter (B-value) rather than a proportion correct scale (P-value) as with classical statistics. For MCQ items, the mean item response time increases as item difficulty increases, but for items that are very difficult (bvalue >= 1.0), the mean response time tends to trail off to less than a minute. This pattern could indicate a "guess and move on" strategy for some examinees who find the items too difficult, but more likely this pattern reflects the behavior of high ability examinees who are able to work very quickly, since items are generally targeted well to an examinee's average ability. Note that for the FBS and MR items, there appears to be a relationship between item difficulty and response time, although it is difficult to see because of the small number of FBS items and the clustering of the MR items in the region of higher item difficulties.

The FBC items appear to take much longer on average than the standard MCQ items but similar to MCQ calculation items. For the RN examination, calculation items tend to favor the lower or easier end of the item difficulty scale.[6]

Summary of Item Fit Measures for the Rasch Model

WINSTEPS is the software program used to calibrate the NCLEX examination items using the Rasch Model (Linacre, 2004). Two measures of model-data fit and their "standardized" transformations are computed as part of the calibration process (Douglas, 1982; Smith & Hedges, 1982). These measures, called "Infit" and "Outfit" have their drawbacks (Smith & Hedges, 1982; Bond & Fox, 2001; Wright & Stone, 1979). While it is true that the issue of model-data fit cannot be evaluated solely in terms of one simple index, these infit/outfit measures can at least provide some rough guide to model-data fit for the Rasch calibration and scaling sequence. The index called "Infit" is considered an "inlier-sensitive" or "information-weighted" fit and is more sensitive to items that are well targeted to a person's ability or pattern. The index called "Outfit" is outlier sensitive and is more sensitive to items with difficulties that are not well targeted on a person or are far from a person's ability estimate.[7]

Table 2 shows the proportion of items flagged for "misfit" for each of the two datasets by item type. Across all two

[6]Note that for the NCLEX-PN examination the item difficulties for FBC items tend to be distributed across the entire difficulty scale

[7]Fit statistics (Infit and Outfit) provide information as to whether the items are functioning as expected and generally meet the assumptions of the Rasch measurement model. In general, those items with fit statistics greater than 2 (criterion of zstd > abs (2) could be considered as not fitting "misfitting" the Rasch model and may need further evaluation in order to include the items in an operational item pool. For more information about the Rasch model the reader should reference Best Test Design by Wright and Stone.

TABLE 2. Summary of Items Flagged for Inft / Outfit by Item Type

		# items	Infit zstd > abs (2)	Outfit zstd > abs (2)	% flagged Infit	% flagged Outfit
RN Items, DS 1	MC	921	139	190	15.1%	20.6%
	FBC	89	0	0	0.0%	0.0%
	FBS	8	0	1	0.0%	12.5%
	MR	88	8	14	9.1%	15.9%
RN Items, DS 2	MC	116	33	37	28.4%	31.9%
	FBC	12	1	1	8.3%	8.3%
	MR	16	0	1	0.0%	6.3%

datasets, the MCQ items flagged for "misfit" based on the Infit index ranged from 15.1% to 28.4%, and the MC items flagged for misfit based on the Outfit index ranged from 20.6% to 32.2%. It should be noted that these percentages of items not fitting the Rasch model ("misfit") are comparable to the percentages of MCQ pretest items.

For the innovative item types, however, Table 2 shows that the range of innovative items flagged for misfit based on the Infit index are 0.0% to 9.8%, and the range of innovative items flagged for misfit based on the Outfit index are 0.0% to 15.9%. This would suggest that the innovative item types generally fit the model as well or at times even better than the MC item types. One explanation for better fit may be related to the reduction of the "guessing space" for the innovative items compared to the traditional MCQ items. For MCQ items, there are only four possible choices, and guessing contributes to the "noise," potentially at probability of 0.25 on average for 4-option MCQs. For MR, FBC, and even FBS items, however, the combination of potential response sets is extremely large so as to reduce the influence of the "noise" due to guessing.

Summary and Conclusions

Based on the two datasets examined, it appears that, in general, the innovative item types may offer measurement properties that are comparable to, and at times better than, the standard MCQ items. Model-data fit for some of the innovative item types appears to be as good as or better than model-fit for the MCQ item types. This property may be related to the decrease, or for practical purposes, elimination, of the guessing noise for some of these items compared to the MCQ items.

Multiple Response
The distribution of item difficulties for MR items is generally good; however, the MR item types have so far tended to show some concentration at the more difficult end of the scale. This could be an artifact of the content chosen for these items and might be resolved with further item writing and focus on the middle and less difficult end of the scale.

Fill in the Blank Calculation
The FBC items on the whole appear to offer much higher item discrimination than the MCQ items, which may generally be considered a positive feature of the items. Another feature of this item type is the tendency to cluster at the mid to lower ends of the item difficulty scale for this RN dataset. These two features of the FBC item types may be a function of the content characteristics of these item samples rather than a general property of the FBC items. However, more study is needed to assess the generalizability of these conclusions about FBC items. For the FBC items, the response time is similar to the MCQ calculation items.

Fill in the Blank Sequence
Although these data have included only handfuls of the FBS item types, it would appear that these items are comparable in item discrimination to the MCQ item types, and span the entire range of item difficulty. The FBS items tend to take the examinee about twice as long as the typical MCQ items. Regarding the FBS items, improvements in the software used to administer these items and using a drag-and-drop format may improve response time and the distribution of item difficulty. (It should be noted that the FBS items have been revised to a drag-and-drop interface in order to address the response time issue.) Nevertheless, item response time needs to be considered in light of overall testing time and test design, particularly for maximum length test-takers.

Based on the data presented in this article, there is comparability between the item characteristics of the innovative items and MCQs. However, the number of FBC and FBS items was relatively small so further investigation will determine if these results can be generalized. The introduction of innovative items does not seem to have impacted NCLEX pass rates. Also, there has not been a

significant increase in the number of candidates running out of time on the exam (NCSBN, 2006). Additionally, innovative items are being introduced into the examination in a measured way. As new item formats are introduced, further investigation will be needed to ensure that the items offer measurement properties that are comparable to if not better than multiple-choice items.

References

Ackerman, P. (2003). Cognitive ability and non-ability trait determinants of expertise. *Educational Researcher*, 32(8), 15-20.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, 17, 283-296.

Bond, Trevor & Fox, Christine (2001). Applying the rasch model. *Fundamental measurement in the human sciences.* Chapter 12, The question of model fit. Mahwah, NJ: Erlbaum.

Douglas, Graham (1982). Issues in the Fit of Data to Psychometric Models. *Education Research and Perspectives* 9:1, 32-43.

Haladyna, Thomas M. (1997). *Writing test items to evaluate higher order thinking.* Allyn & Bacon. Needham Heights, MA.

Jodoin, M. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1-15.

Linacre, J.M. (2003). *A users guide to WINSTEPS: Rasch measurement computer program*, Chicago, IL: MESA Press.

Linacre, J. M. (2004). *WINSTEPS Rasch Measurement.* Version 3.50 (February, 2004). Chicago.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Martinez, M. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 218-241.

McDonald, M. (2002). *Systematic assessment of learning outcomes: Developing multiple- choice exams*. Jones & Bartlett. Sudbury: MA

McHenry, J. & Schmitt, N. (1994). Chapter 12: Multimedia Testing in *Personnel Selection and Classification.* J. Harris, M. Rumsey, & Walker, C. (Eds.). Erlbaum associates. Hillsdale: NJ.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary models.* Chicago: Scientific Software.

National Council of State Boards of Nursing (NCSBN). (2006). 2006 Annual NCSBN Business Book. Chicago. NCSBN.

Parshall, C., Davey, T., & Pashley, P. (2000) Innovative item types for computerized testing in *Computerized Adaptive Testing Theory and Practice*. Van der Linden, W. & Glas, C. (Ed.) Kluwer Academic Publishers. Netherlands.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institute; reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press.

Smith, R & Hedges, L. (1982). Comparison of likelihood ratio χ^2 and pearsonian χ^2 tests of fit in the rasch model. *Education Research and Perspectives* 9:1, 44-54.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.