Graduate Management Admission Council®

Developing Item Variants: An Empirical Study

Anne Wendt
National Council of State Boards of Nursing
Shu-chuan Kao and Jerry Gorham
Pearson VUE
Ada Woo
National Council of State Boards of Nursing

Presented at the Item and Pool Development Paper Session, June 3, 2009

Abstract

Large-scale standardized test have been widely used for educational and licensure testing. In computerized adaptive testing (CAT), one of the practical concerns for maintaining large-scale assessments is to ensure adequate numbers of high-quality items that are required for item pool functioning. Developing items at specific difficulty levels and for certain areas of test plans is a well-known challenge. The purpose of this study was to investigate strategies for varying items that can effectively generate items at targeted difficulty levels and specific test plan areas. Each variant item generation model was developed by decomposing selected source items possessing ideal measurement properties and targeting the desirable content domains. 341 variant items were generated from 72 source items. Data were collected from six pretest periods. Items were calibrated using the Rasch model. Initial results indicate that variant items showed desirable measurement properties. Additionally, compared to an average of approximately 60% of the items passing pretest criteria, an average of 84% of the variant items passed the pretest criteria.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Wendt, A., Kao, S., Gorham, J., & Woo, A. (2009). Developing item variants: An empirical study In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Ada Woo, National Council of State Boards of Nursing (NCSBN), 111 E. Wacker Drive, Suite 2900, Chicago, IL 60601 U.S.A. Email: awoo@ncsbn.org

Developing Item Variants: An Empirical Study

Large-scale standardized tests have been widely used for educational and licensure testing. In computerized adaptive testing (CAT), one of the practical concerns for maintaining large-scale assessments is ensuring the availability of adequate numbers of high-quality items that are required for item pool functioning. Developing items at specific difficulty levels and for certain areas of test plans is a well-known challenge. The purpose of this study was to investigate strategies for effectively generating items at targeted difficulty levels and specific test plan areas.

Theoretical Background

Earlier researchers (LaDuca, Staples, Templeton, & Holzman, 1986, Bejar, 1996) described item modeling as a construct-driven approach to test development that is potentially validity-enhancing. Earlier research focused on mirroring cognitive processes in answering surveys for psychological performance (Bejar, 1993; Embretson & Gorin, 2001; Embretson, 1999; Bejar & Yocom, 1991), with the intention of generating isomorphic items. For large-scale testing, some item models are more statistics-driven (e.g., Glas & van der Linden, 2003) and others are more content-driven (e.g., Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003). Each item model provides templates that allow the decomposition of knowledge or skills and identification of the key components that constitute meaningful new items.

As described by Shye, Elizur and Hoffman (1994), item features can be mapped into an item by a set of rules using Guttman's (1969) facet theory. That is, by identifying the fixed and variable elements in items, stimulus features are substitutable in the variable elements for generating structurally equivalent items. In this study, variant item models were developed by decomposing the selected source items possessing ideal measurement properties and targeting the desirable content domains. The selected source items were operational items in a CAT examination for nurses and were used to set up the basic frame of the new items. That is, the sentence structure in source items was fixed. Item length and grammatical syntax were also fixed. Variant items can be defined as generated items from a model in which specific item stimulus features can vary. As Table 1 shows, four item models were proposed to generate item variants. Ideally, the proposed models would generate variant items with similar item difficulty and other psychometric features.

Method

Data

All variant items were administered as pretest items to at least 400 candidates in order to gather statistical information. No more than three variant items generated from the same source item were selected for one pretest pool, and the administration of the pretest items was controlled through a masking process. This strategy was used in order to reduce the possibility of administering similar items to the same candidate so that the candidate would not think the same item was administered twice.

Procedures

In this study variant items were evaluated for both classical test theory (CTT) and item response theory (IRT) properties using the exam data of first-time U.S.-educated candidates. Even though the CTT item statistics are sample dependent, comparisons are still legitimate when

item statistics are from random samples. The random selection scheme implemented in the test driver ensured that candidates were exposed to items randomly sampled from the pretest pool. On the other hand, IRT parameters were calibrated using Winsteps, a Rasch-based computer program (Linacre & Wright, 2001).

Table 1. Variant Item Models

Item Model	Definition in Item Development
Key	Delete original key and replace it with a new key
Stem	Change stimulus in stem
Distractor	Delete one original distractor and add a new distractor
Other	Add key
	Add key and extra distractor
	Add key and change stem
	Change key and distractor
	Change stem and distractor
	Change stem and key
	Change key, stem, and distractor

In this report, the summary statistics of item *p*-value (proportion correct) difference, item-θ point-biserial correlation (r_{pb}) difference, item response time difference, and item difficulty difference were provided. Due to the ideal scale properties provided by the Rasch model, the IRT item property (item difficulty difference) was explored with greater detail: the group-level comparison was conducted using analysis of variance (ANOVA) to ascertain whether the item *b*-value (difficulty) difference shifted more for one group than for others; the item-level comparisons were performed using *t*-tests to determine whether the item *b*-value difference was significantly greater than calibration error.

Results

Using the proposed item models, the 341 variant items generated from 72 source items had enough item exposures (greater than or equal to 400) for analysis. The average item exposure of the items in the analysis was 556. As shown in Table 2, the percent of items passing pretesting was satisfactory. The pretest passing rates varied from 78.85% to 100.00%. Compared to an average of approximately 60% of the items passing pretest, 84% of the variant items passed the statistical pretest criteria.

The summary statistics of the difference of item *p*-values between the source and that from variant items are reported in Table 3. The *p*-value difference was calculated by subtracting the source item *p*-value from the variant item *p*-value. As indicated in Table 3, the means of the *p*-value difference for different models were similar, varying from –0.001 to 0.066. However, the "Other" model exhibited a relatively large SD for *p*-value difference, indicating relatively huge discrepancies. Concerning item type, MR items had a lower average item *p*-value than their source item by –0.240, indicating that variant items tended to be more difficult. For all 341 variant items in the analysis, the mean *p*-value difference was 0.016 with a SD of 0.186.

Table 2. Pretest Results of Variant Items

Exam	No. of Items in Pretest	No. of items in Analysis (Exposure > 400)	No. of Items Passed Pretest	Percentage of Passing*
Pretest pool 1	147	93	74	79.57%
Pretest pool 2	199	104	82	78.85%
Pretest pool 3	21	20	18	90.00%
Pretest pool 4	61	59	53	89.83%
Pretest pool 5	31	31	31	100.00%
Pretest pool 6	34	34	27	79.41%
Total	493	341	285	83.58%

*(Number of items passed pretest) / (Number of items in analysis).

Table 3. Summary Statistics of Item *p*-Value Difference

Factors	N	Mean	SD	Minimum	Maximum
Item model					
Key	71	–0.001	0.190	–0.465	0.330
Stem	105	0.039	0.159	–0.661	0.445
Distractor	101	0.066	0.124	–0.271	0.425
Other	64	–0.080	**0.257**	–0.613	0.443
Item type					
FC	39	0.063	0.092	–0.111	0.289
MC	269	0.041	0.169	–0.661	0.445
MR	33	**–0.240**	0.204	–0.613	0.099
Total	341	0.016	0.186	–0.661	0.445

Table 4 reports the summary statistics the difference between the item-θ point-biserial correlation of the source and that of the variant items. This type of point-biserial correlation reflects the association between the item scores (0 = incorrect, 1 = correct) and the CAT final θ estimates. The difference of the item-θ point-biserial correlation was calculated by r_{pb}(diff) = variant item r_{pb} – source item r_{pb}

As Table 4 shows, the means of r_{pb}(diff) for item models varied from –0.034 to –0.062. The means of r_{pb}(diff) for item type were similar, ranging from –.053 to 0.007. Among three item types, FC items had the smallest mean r_{pb}(diff) of 0.007 with the smallest SD of 0.047, indicating that FC variant items had stable item discrimination power. Overall, the item-θ point-biserial correlation difference had a mean of –0.045 and a SD of 0.083.

Table 4. Summary Statistics of r_{pb}(diff)

Factors	*N*	Mean	SD	Minimum	Maximum
Item model					
Key	71	-0.062	0.095	-0.265	0.150
Stem	105	-0.034	0.074	-0.280	0.138
Distractor	101	-0.045	0.079	-0.232	0.164
Other	64	-0.043	0.087	-0.348	0.086
Item type					
FC	39	0.007	0.047	-0.091	0.109
MC	269	-0.053	0.087	-0.348	0.164
MR	33	-0.038	0.052	-0.162	0.067
Total	341	-0.045	0.083	-0.348	0.164

The summary statistics in Table 5 describe the difference between response time of the source and that of the variant items. The difference of the item response time was calculated as variant item response time minus source item response time. The means of the response time difference for item model varied from –1.546 to 10.431 seconds. The means of the response time difference for item type varied substantially. FC items had a large mean response time difference of 31.024 seconds with a SD of 48.635 seconds, indicating that the time required to process nursing computation was considerably different among candidates. On the contrary, for MC items the response item difference was only –1.160 with a SD of 8.571. Overall, the item response item difference had a mean of 4.775 seconds and a SD of 21.650 seconds.

Table 5. Summary Statistics of Response Time Difference in Seconds

Factors	N	Mean	SD	Minimum	Maximum
Item model					
Key	71	0.478	7.952	–21.622	20.350
Stem	105	**10.431**	33.845	–79.131	95.081
Distractor	101	–1.546	8.785	–30.178	14.002
Other	64	**10.236**	16.658	–23.249	46.051
Item type					
FC	39	**31.024**	**48.635**	–79.131	95.081
MC	269	–1.160	8.571	–30.178	20.350
MR	33	22.132	10.472	4.441	46.051
Total	341	4.775	21.650	–79.131	95.081

In Table 6, the summary statistics of the difference between the item *b*-value of the source and that of the variant items are reported. The item *b*-value was calibrated from the Rasch model, which estimates item difficulty on the logit scale. The difference of *b*-values was calculated by the formula *b*(diff) = variant item *b*-value – source item *b*-value.

The means of *b*(diff) for item model varied from –0.013 to 0.589. The means of *b*(diff) for item type varied substantially. MR items had a large mean *b*(diff) of 1.518 with a large SD 1.201. MC items had a small mean *b*(diff) of 0.056 with a SD of 0.840. For all 341 items, *b*(diff) had a mean of 0.173 and a SD of 0.955.

Table 6. Summary Statistics of *b*(diff)

Factors	*N*	Mean	SD	Minimum	Maximum
Item model					
Key	71	0.291	0.905	-1.698	2.559
Stem	105	0.017	0.776	-2.205	3.672
Distractor	101	-0.013	0.631	-2.388	1.632
Other	64	0.589	1.448	-2.421	5.013
Item type					
FC	39	-0.165	0.422	-1.079	0.552
MC	269	0.056	0.840	-2.421	3.672
MR	33	**1.518**	1.201	-0.025	5.013
Total	341	0.173	0.955	-2.421	5.013

For the convenience of describing the distribution of *b*(diff) for item model, item type, and content area, Figures 1 to 3 provide box plots for each subgroups of interest.

Figure 1. Box Plot of *b*(diff) for the Variant Item Model

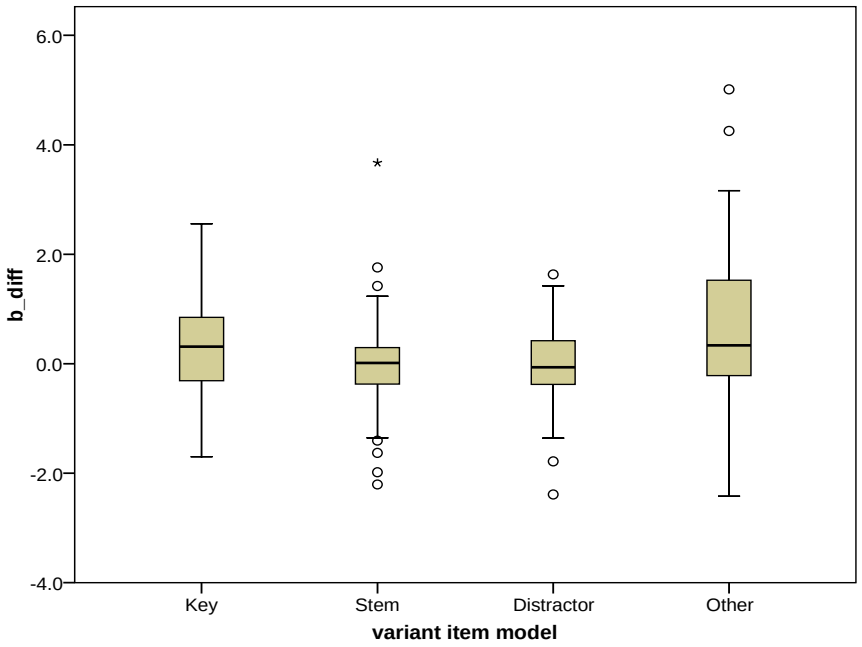


Figure 2. Box Plot of *b*(diff) for the Variant Item Type

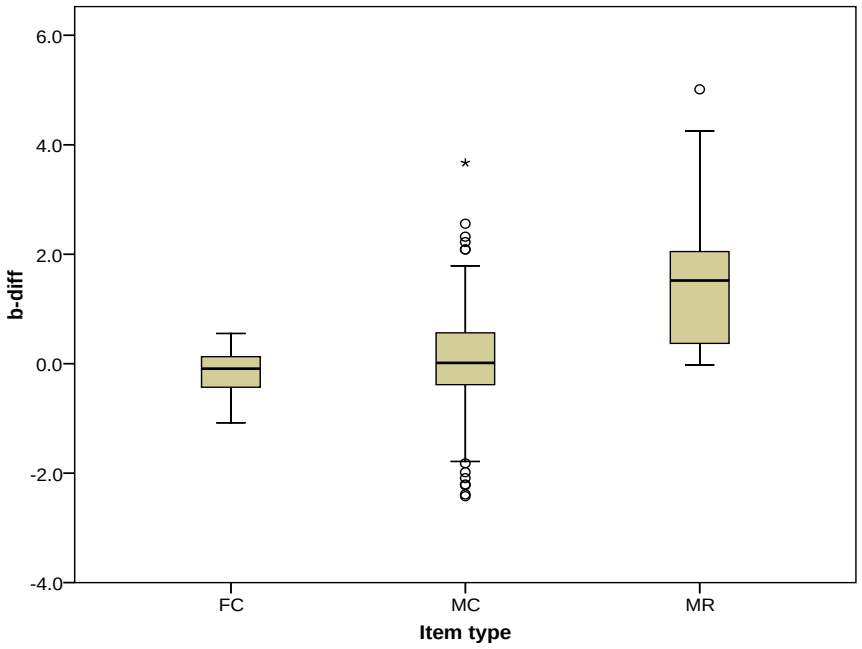
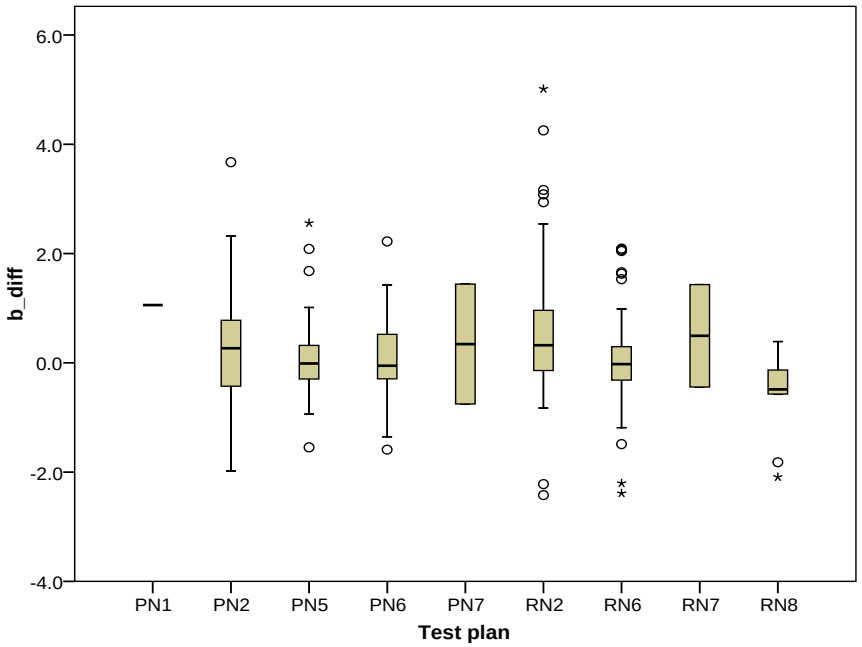


Figure 3. Box Plot of *b*(diff) for the Test Plan



In order to test the group-level *b*-value difference, a two-way ANOVA was performed, using the factors item type and variant model. The ideal design would be to include the test plan in the three-way ANOVA model, but the resulting empty cells in the joint distribution could cause serious problems in the significant test.

First, Levene's homogeneity test was significant ($F_{(7, 333)} = 5.519$, $p < .05$), indicating that the variances in the different groups of the 4 (item model) × 3 (item type) design were not homogeneous. According to Lindman (1974, p. 33) and Box (1954), the F statistic is quite robust against the violations of the homogeneity assumption. The F test can provide information concerning the group mean difference but special caution should be paid in interpreting the results.

The ANOVA model in Table 7 was significant ($F_{(7, 333)} = 15.681$, $p < .05$), indicating that at least one group mean was significantly different from others. Given that the interaction of item model and item type was not significant ($F_{(2, 333)} = 0.203$, $p > .05$), it was appropriate to explore the main effects for item model and item type. The main effect for item model was significant ($F_{(2, 333)} = 8.379$, $p < .05$) with an effect size of 0.067. The main effect of item type was also significant ($F_{(3, 333)} = 2.644$, $p < .05$) with a effect size of 0.033.

Table 7. Summary Results From a Two-Way ANOVA

Source	SS	df	MS	*F*	Sig.	Partial η^2
Corrected model	76.811(a)	7	10.973	15.681		.248
Intercept	5.945	1	5.945	8.496	.004	.025
Item model	**16.758**	**2**	**8.379**	**11.974**	**.000**	**.067**
Item type	**7.931**	**3**	**2.644**	**3.778**	**.011**	**.033**
Item model × Item type	.284	2	.142	.203	.816	.001
Error	233.023	333	.700			
Total	319.992	341				
Corrected total	309.834	340				

Note. $R^2 = .248$ (adjusted $R^2 = .232$).

In order to identify which group means were different from others, Bonferroni's post-hoc comparison was conducted for factor variant model and item type, respectively. Tables 8 and 9 tabulate all possible paired comparisons for item model and item type, respectively. Concerning item model, the "Other" model seemed to generate harder items more often than the Stem and Distractor models. With regard to item type, MR variant items tended to have a positive shift on item difficulty more often that the FC and MC variant items. Since the interaction was not significant, it is legitimate to conclude that items generated from the "Other" model with the item type of MR tended to have a more noticeable positive shift on item difficulty than the rest of the variant items.

Table 8. Multiple Comparisons for Different Item Models

(I) Item Model	(J) Item Model	Mean Difference (I-J)	Std. Error	Sig.	95% CI Upper Bound	95% CI Lower Bound
Key	Stem	0.274	0.129	0.204	–0.067	0.615
	Distractor	0.303	0.130	0.119	–0.040	0.647
	Other	–0.298	0.144	0.236	–0.681	0.084
Stem	Key	–0.274	0.129	0.204	–0.615	0.067
	Distractor	0.030	0.117	1.000	–0.280	0.339
	Other	**–0.572***	**0.133**	**0.000**	–0.924	-0.220
Distractor	Key	–0.303	0.130	0.119	–0.647	0.040
	Stem	–0.030	0.117	1.000	–0.339	0.280
	Other	**–0.602***	**0.134**	**0.000**	–0.956	-0.247
Other	Key	0.298	0.144	0.236	–0.084	0.681
	Stem	**0.572***	**0.133**	**0.000**	0.220	0.924
	Distractor	**0.602***	**0.134**	**0.000**	0.247	0.956

*The mean difference was significant at the .05 level.

Table 9. Multiple Comparisons for Item Types

(I) Item Type	(J) Item Type	Mean Difference (I-J)	Std. Error	Sig.	95% CI Upper Bound	95% CI Lower Bound
FC	MC	–0.221	0.143	0.373	–0.566	0.124
	MR	**–1.683***	**0.198**	**0.000**	–2.159	–1.207
MC	FC	0.221	0.143	0.373	–0.124	0.566
	MR	**–1.462***	**0.154**	**0.000**	–1.833	–1.091
MR	FC	**1.683***	**0.198**	**0.000**	1.207	2.159
	MC	**1.462***	**0.154**	**0.000**	1.091	1.833

* The mean difference was significant at the .05 level.

In addition to comparing the group-level *b*-value differences, item-level analysis was conducted to explore whether the *b*-value differences were significantly different from random error in test calibration. The decision of significant difference was determined by calculating the 95% confidence interval using the calibration standard error. As shown in Table 10, the "Stem" model generated the highest percentage of items (37.14 %) that had different *b*-values from the source items. Concerning item type, FC had the highest percentage of items (43.59 %) that exhibited significantly different *b*-values. Overall, 23.17% of variant items had significant *b*-value changes, indicating truly easier or more difficult items.

Table 10. Item-Level Significant Tests of *b*-Value Differences

Item Model	Item Type	Sig. Yes	Sig. No	Total
Key	MC	0	1	1
	MC	7	32	39
	MC	2	8	10
	MC	0	11	11
	MC	0	2	2
	MC	1	6	7
	MC	1	0	1
Stem	MC	2	21	23
	MC	8	6	14
	MC	6	9	15
	MC	4	4	8
	FC	15	18	33
	MC	4	5	9
	MC	0	3	3
Distractor	MC	2	25	27
	MC	7	16	23
	MC	6	19	25
	MC	2	11	13
	MC	1	6	7
	MR	0	1	1
	MC	0	1	1
	MC	2	2	4
Other	MC	1	1	2
	MC	2	2	4
	MC	1	0	1
	MC	0	7	7
	MR	2	24	26
	FC	2	4	6
	MC	1	6	7
	MR	0	6	6
	MC	0	1	1
	MC	0	4	4
Key Total		11	60	71
Stem Total		39	66	105
Distractor Total		20	81	101
Other Total		9	55	64
		0	1	1
		12	79	79
		19	32	32
		13	39	39
		0	2	2
		9	52	52
		24	46	46

Table 10, continued
Item-Level Significant Tests of *b*-Value Differences

Item Model	Item Type	Sig. Yes	Sig. No	Total
		0	2	2
		2	9	9
	FC Total	17	22	39
	MC Total	60	209	269
	MR Total	2	31	33
Grand Total		79	262	341

Note:. the significant difference level was $\alpha = .05$.

Conclusions

The purpose of this study was to propose item models that can effectively generate items at targeted difficulty levels and specific test plan areas. Based on the results from this study, the four proposed item models can generate item variants with higher passing rate for pretests based on the statistical criteria. Concerning the group-level item difficulty shift, the "Other" model with the MR item type had the largest *b*-value increase. At the item level, 76.83% variant items did not have a significantly different item difficulty shift. These items can be considered to be isomorphic items of the item variant source. In summary, the proposed variant item models offer an efficient way to develop items possessing desirable psychometric characteristics.

Other than generating items for CAT pools, the proposed models can be further expanded to investigate the construct validity for the current licensure exams. As Embretson (1983) indicated, manipulating item stimulus features directly impacts one aspect of construct validity or the construct representation. The identified components, along with the associated item stimulus features, are the basis for the item specifications. The knowledge and cognitive abilities involved in the item-solving process can be better identified and verified if the proposed models can be expanded to incorporate cognitive theories and can be continued in the future.

Another possible direction to expand this research is the explanation of the association of item similarity and the change of item psychometric properties. Whichever item model is employed, multiple variant items can be generated from a single source item, and the "family-wise" item statistics shift can be an interesting topic for investigation. By comparing the semantic content of test items, the functioning of item models can be better explained and supported.

It is expected that the use of variant item models can make item development more cost-efficient and less labor-intensive. Most importantly, the characteristics of the new items seem to be better controlled and more predictable than the "standard" methods for developing items (item writing and item review). Though this research was based on specific licensure exams, the methodology of this study might be applicable to other testing programs.

References

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium*. Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1993). Testing reading comprehension skills: Part 2. Getting students to talk about taking a reading test (A pilot study). *Reading in a Foreign Language*, *6*, 425-438.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2*(3), 3-28.

Bejar, I. I. & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement. 15*, 2129-137.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: II Effect on inequality of variance and correlation of errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484-498.

Embretson, S. E. (1983). Construct validity: construction representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407-433.

Embretson, S. E. & Gorin J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343-368.

Glas C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247-261.

Guttman, L. (1969). *Integration of test design and analysis: Proceedings of the 1969 invitational conference on testing problems*. Princeton, NJ: Educational Testing Service.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modeling procedure for constructing content-equivalent multiple-choice questions. *Medical Education*, *20*(1), 52-56.

Linacre, J. M. & Wright, B. D. (2001). *A user's guide to WINSTEPS: Rasch model computer program*. Chicago: MESA Press.

Lindman, H. R. (1974). *Analysis of variance in complex experimental designs*. New York, NY: W. H. Freeman.

Shye, S., Elizur, D. & Hoffman, M (1994). *Introduction to facet theory*. Thousand Oaks, CA: Sage.