Developing Item Variants: An Empirical Study

Anne Wendt

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®.

Copyright © 2009 by the Authors

All rights reserved. Permission is granted for non-commercial use.

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

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.

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

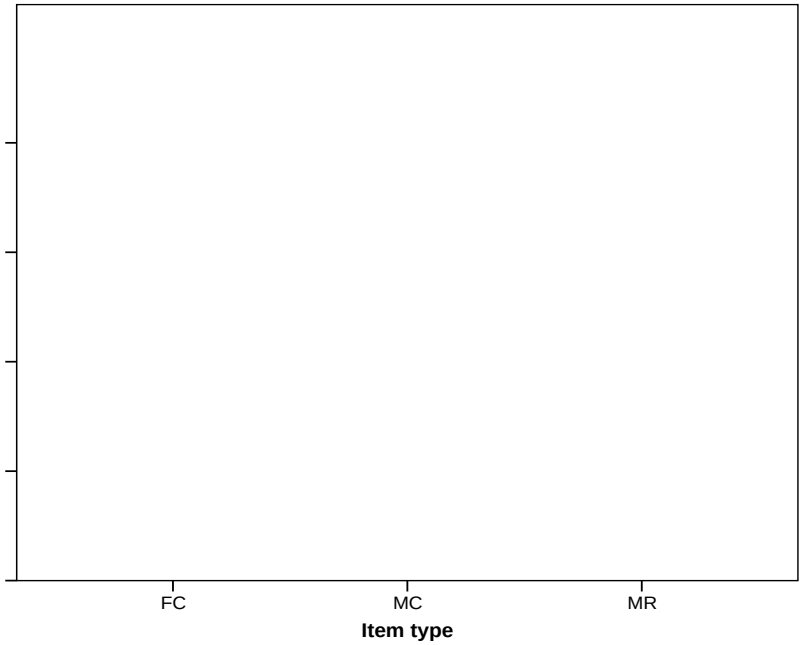
Table 4 reports the summary statistics the difference between the item-*T* point-biserial correlation of the source and that of the variant items. This type of point-biserial correlation reflects the association between the item scores (0 = incorrect, 1 = correct) and the CAT final *T* estimates. The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

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

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

The means of

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



First, Levene's homogeneity test was significant ($_{(7, 333)} = 5.519$, $< .05$), indicating that the variances in the different groups of the 4 (item model) u3 (item type) design were not homogeneous. According to Lindman (1974, p. 33) and Box (1954), the statistic is quite robust against the violations of the homogeneity assumption. The 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 ($_{(7, 333)} = 15.681$, $< .05$), 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 ($_{(2, 333)} = 0.203$, $> .05$), it was appropriate to explore the main effects for item model and item type. The main effect for item model was significant ($_{(2, 333)} = 8.379$, $< .05$) with an effect size of 0.067. The main effect of item type was also significant ($_{(3, 333)} = 2.644$, $< .05$) with a effect size of 0.033.

Table 7. Summary Results From a Two-Way ANOVA

Source	SS	df	MS		Sig.	Partial ^{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 uItem type	.284	2	.142	.203	.816	.001
Error	233.023	333	.700			
Total	319.992	341				
Corrected total	309.834	340				

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

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 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 ofb-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 . = .05.

Conclusions

References

Bejar, I. I. (1996).
. 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). , , 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.
(3), 3-28.

Bejar, I. I. & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. , 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. , 484-498.

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

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

Embretson, S. E. & Gorin J. S. (2001). Improving construct validity with cognitive psychology principles. , , 343-368.

Glas C. A. W. & van.lc28E. Pitv.MC /P <</MCID 9 >-9.P <</MCID(Journal A: Tc -0.002hT0 a 9.085 0 &65 Tc