CLEAR Exam Review

Volume XX, Number 2
Summer 2009

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@clearhq.org, 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.

CLEAR Exam Review

VOLUME XX, NUMBER 2 SUMMER 2009

Contents

COLUMNS

ARTICLES

Developing Models That Impact Item Development

ANNE WENDT, PhD, RN, CAE
National Council of State Boards of Nursing (NCSBN)

SHU-CHUAN KAO
Pearson VUE

JERRY GORHAM
Pearson VUE

ADA WOO
National Council of State Boards of Nursing (NCSBN)

Introduction

One of the most important practical concerns for licensure and certification programs is ensuring adequate numbers of high quality items for item banks. Item development is expensive and developing items at specific difficulty levels and for certain areas of test plans are additional challenges. A strategy that can assist licensure and certification programs to efficiently and effectively develop needed items for targeted areas can produce significant benefits.

Several years ago, the National Council of State Boards of Nursing (NCSBN®) began investigating models for varying key elements of an item in order to produce items with desirable measurement properties in needed test plan categories. Previous research found that various item models can be developed by using key elements from selected "source" items possessing ideal measurement properties (Bejar, Lawless, Morley, Wagner, Bennett, and Revuelta, 2003 and Glas, C & van der Linden W. 2003). The selected "source" items reported in this study are operational items in a computerized adaptive examination for nurses. The source items are used to construct a basic template for new variant items. Variant items can be defined as generated items from a model in which specific item stimulus features can vary. As Table 1 illustrates, four item models were used to generate item variants. The major elements of the models that were varied included the stem, the key, the distracter and multiple changes to an item.

Figure 1 provides an example of the multiple-choice source item and Figures 2 through 5 provide examples of the item models with the elements that were varied shown underlined.

In this study, 72 source items were varied, administered to over 400 candidates and analyzed. The 72 items produced 341 variant items that met pretest criteria in targeted areas of the test plan. The percent of items passing pretest was quite satisfactory. The pretest passing rates from different pretest pools varied from 79% to 100% with an average of 84% of variant items meeting statistical criteria. The 84% pass rate is quite good when compared to an average of approximately 60% for the operational program. Additionally, the item variants developed using these four models addressed targeted areas of content and item difficulty. Approximately 53% of the items (N=181) were similar to the source in terms of item difficulty—varying less than 0.05 logit.

Variant items were evaluated using both Classical Test Theory (CTT) and Item Response Theory (IRT) properties. The summary statistics of item p-value difference, item-theta point-biserial correlation (rpb) difference, item response time difference, and IRT item difficulty difference were calculated but only the information on p-values and response times are reported within this article for the sake of brevity.

The summary statistics of the difference of item p-values between the source and the variant items are reported in Table 2. The p-value difference is calculated by subtracting the source item p-value from the variant item p-value (See Equation 1). Thus, a positive p-value difference would indicate that the variant item was easier than the source item. As indicated in Table 2, the means of the

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
Distracter	Delete one original distracter and add a new distracter
Other	Add key
	Add key and extra distracter
	Add key and change stem
	Change key and distracter
	Change stem and distracter
	Change stem and key
	Change stem, key and distracter

A client is receiving a 500-mg sodium diet. Which of the following foods should the nurse suggest when assisting this client to select a daily menu?

1. Whole wheat bread, macaroni salad and skim milk
2. Cottage cheese and fresh fish
3. Spinach and celery
4. Chicken, puffed wheat and fresh fruit* (key)

FIGURE 1. Source Item: Multiple-Choice (MC)

A 70 year-old client has been prescribed a low sodium diet. Which of the following foods should the nurse recommend?

1. Whole wheat bread, macaroni salad and skim milk
2. Cottage cheese and fresh fish
3. Spinach and celery
4. Chicken, puffed wheat and fresh fruit* (key)

FIGURE 2. Item Model: Stem Varied

A client is receiving a 500-mg sodium diet. Which of the following foods should the nurse suggest when assisting this client to select a daily menu?

1. Whole wheat bread, macaroni salad and skim milk
2. Cottage cheese and fresh fish
3. Spinach and celery
4. Baked chicken, rice and fresh fruit* (key)

FIGURE 3. Item Model: Key Varied

A client is receiving a 500-mg sodium diet. Which of the following foods should the nurse suggest when assisting this client to select a daily menu?

1. Carrots and mustard greens
2. Cottage cheese and fresh fish
3. Spinach and celery
4. Chicken, puffed wheat and fresh fruit* (key)

FIGURE 4. Item Model: Distracter Varied

A client is receiving a 500-mg sodium diet. Which of the following foods should the nurse suggest when assisting this client to select a daily menu? **Select all that apply.**

1. Macaroni salad and skim milk
2. Rice and baked fish* (key)
3. Spinach and celery
4. Chicken and fresh fruit* (key)
5. Artichokes and beets

FIGURE 5. Item Model: Other Multiple Response (MR)

Table 2. 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
Distracter	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

p-value difference of the four models are quite similar. The "Other" model, however, exhibits a relatively large standard deviation (SD) for p-value difference, indicating relatively large discrepancies in item difficulty. This is not surprising since the "Other" model is not a pure model. Rather, the "Other" model consists of multiple models requiring multiple changes to an item ranging from adding a key to a combination of adding a key, distracter and/or revising the stem. Also, as shown in Table 2, concerning item type, multiple response (MR) items have a lower average item p-value than their source item by -0.240 which indicates that variant MR items tend to be more difficult than the source multiple-choice (MC) items. This finding that MR items tend to be more difficult than a comparable MC item is consistent with prior research (Wendt, 2008). On average for all of the 341 items, there were minimal changes in the item difficulty (p-value) when one of the four models shown in Table 1 is used to vary the source item.

$$P_{Difference} = P_{Variant} - P_{Source}$$

The summary statistics in Table 3 describe the difference between the response time of the source and that of the variant items. The difference of the item response time is calculated by subtracting the source item response time from the variant item response time (See Equation 2). Thus, a positive number would indicate a longer response time for the variant item. The means of the response time difference for item model varies from -1.546 to 10.431 seconds. Examinees take about 10 seconds longer to respond to the items that were varied by changing the stem and changing multiple elements of the item. The means of the response time difference for item type varies substantially. Examinees responding to the fill-in-the-blank calculation items (FC) took about 31 seconds longer to respond to the variant item as compared to the source item. It should be noted, however, that there was a large variation (SD=48.635 seconds) in the time examinees took to respond which could indicate that the time required to process nursing computations is considerably different among examinees. In general, it is likely that creating an answer as with FC items may take longer than selecting a correct answer as with MC calculation items. However, FC items tend to be more discriminating than the MC calculation items (Wendt, 2008). For MC items, the response item difference between the source item and variant item is about one second faster Overall, the examinees spent about five seconds longer on the variant item as compared to the source item. This slight increase in response time could be attributable to increased time needed for cognitive processing as some of the variants were more difficult. A modest increase in response time should not be a deterrent to developing item variants. Test developers need to be aware of the potential for examinees to spend more time on difficult items and consider this into the total time allocated for testing.

$$T_{Difference} = T_{Variant} - T_{Source}$$

Based on the results from this study, the four proposed item models can effectively generate items at targeted difficulty levels and content domains. Using these models, it is possible to efficiently develop item variants that pass pretest because the variants are developed based on items in the existing operational pools by either NCLEX content staff or volunteer item writers using specific item guidelines. By varying items that have passed pretest, it is expected that the use of variant item models can make item development more cost-efficient and less labor-intensive. The variant items use many of the same references that are used to validate the correct answer and rule out distracters of the source item—which saves time and reduces potential errors. More importantly, the characteristics of the new items seem to be better controlled and more predictable than the "standard" methods for developing items. Items developed with a traditional method are

Table 3. Summary Statistics of 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
Distracter	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

generally reviewed multiple times prior to pretesting. A newly written item may be deemed unsuitable for use in each of these development steps. The current study describes an item development method that allows test developers and volunteers to start with items that have passed pretest and have known statistical characteristic. For example, the test developer would know which distracter is not being selected and create a new distracter. Thus, generating variants from operational source items can provide greater control over the development process. It should be noted that this method of using models to vary items is not meant as a substitute for the traditional item development method, but rather to supplement it. Though this research is based on specific licensure exams, the methodology of this study may be applicable to other licensure and certification programs.

References

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.

Embretson, S. E. & Gorin J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343-368.

Glas C. a. W. & van der Linden. (2003). Computerized Adaptive Testing with item cloning. *Applied Psychological Measurement*, 27(4), 247-261.

Wendt, A. (2008). Investigation of the item characteristics of innovative item formats. *CLEAR Exam Review*. 19 (1) p. 22-27.