

The Care and Feeding of Innovative Items

Kirk A. Becker
Pearson VUE

Abstract

While many of the item types referred to as “innovative items” have been available and in use for some time, they are rarely produced and used in large scale. This paper will present practical considerations for new and existing programs that are beginning to use innovative items. Systems for psychometric analyses and standard feedback for content developers regarding innovative item types is unlikely to be established for most testing programs. Because innovative items are typically just a small portion of a testing program, the analysis and scoring may be treated as a hand-crafted, individualized process. If innovative items are going to be part of a testing program for the long run, this will have to change. By instituting the correct processes in writing, reviewing, analyzing, and scoring innovative items, the time and expense involved can be reduced while increasing item retention.

Introduction

Practical considerations in computer-based testing (Parshall, Spray, Kalohn, & Davey, 2001) defines innovative items as items that “depart from traditional, discrete, text-based, multiple-choice format.” This can include anything from full simulations of medical or IT procedures, but can also include multiple-choice items that use audio, video, or graphical content within stems or options. Even the replacement of radio buttons (select one) with check boxes (select all that apply) in text-based items creates a significantly different item type from both a candidate-level and psychometric perspective.

Innovative items have been available for use in most CBT (and even PBT) platforms for over a decade. Despite this availability, the majority of items administered are plain-text multiple-choice items, and many testing programs use only these items. For a variety of reasons this status-quo has slowly begun to change, and large-scale testing programs have begun to investigate and adopt innovative items. Innovative items have been used on a large scale in nursing, accounting, language-testing, and driving exams. By thinking through and working with the complexity of data structures and scoring modules we have developed recommendations for the use of these items on a large scale basis. The goal of this paper is to present and consolidate data, advice, and information on the operational use (development, analysis, storage, etc.) of innovative items based on experience with these programs.

Item Types

While the term “innovative item” encompasses a wide range of possible test content, there are several common item types that are widely used and available. This section will explore the characteristics of these common item types.

Multiple-Choice Multiple-Response (MCMR)

Description: MCMR items are similar to standard multiple-choice items, however they require test takers to select all options that meet some criteria.

Instructions: MCMR items either specify the number of correct responses or instruct test takers to select all options that apply.

Data considerations: Multiple-select multiple choice offer two options for response analysis. Traditional option analysis (average scores/measures, option-total correlations, frequencies, pvalues, etc.) methods can be applied either to the entire response or the disaggregated response. For example, if candidates select two options out of 5 available options (A-E), the entire response will include “AB”, “AC”, “AD”, “AE”, etc. The disaggregated responses would be five separate columns indicating whether the candidate had chosen each of response A-E. Table 1 demonstrates these two data structures.

Table 1. Example of Aggregate and Disaggregate Responses

Aggregate Response	Disaggregated Response				
	A	B	C	D	E
AB	1	1	0	0	0
AC	1	0	1	0	0
AD	1	0	0	1	0
AE	1	0	0	0	1

Scoring: MCMR items can be scored dichotomously or polytomously. Polytomous scoring provides points for the selection of correct options and may subtract points for selecting incorrect options. If negative scoring is used, the minimum possible score should be set to 0. The decision to penalize incorrect options will depend on the instructions and setup of the MCMR items. When items are structured to limit the number of options that can be selected, partial credit scoring accounts for incorrect options without negative scoring. For example if a test taker can only select two options, selection of an incorrect option reduces the number of possible points by 1. If test takers are not limited in their ability to select responses (e.g., test takers can select as many options as they choose), then either dichotomous scoring or polytomous scoring with negative scoring is required. The subtraction of points based on incorrect choices both moderately improves the discrimination of the items relative to correct only and removes the risk of receiving high scores by selecting all options. When scored dichotomously, multiple-response items may be harder than the average single-response multiple-choice item for a program – providing a tool for increasing the difficulty of an item bank when necessary.

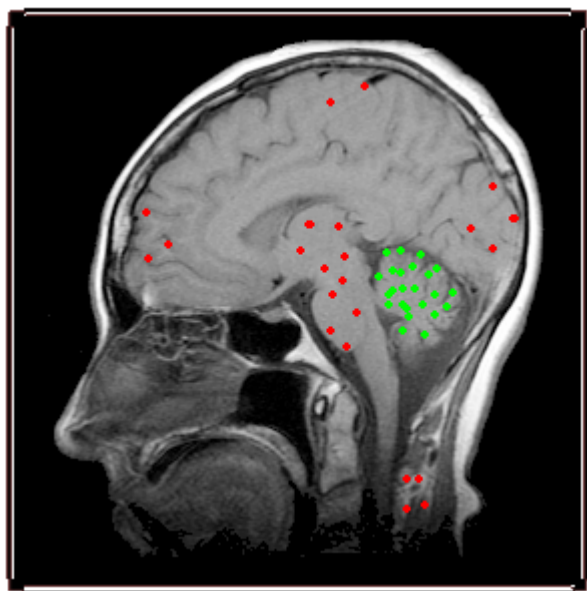
Hotspot

Description: The hotspot item type can be either single-response or multiple-response. A hotspot item requires a candidate to click (or touch) the region of a stimulus that meets the requirement of the stem (e.g., identifying a region on an image produced from an MRI). While hotspot items typically involve images, they can also employ text, requiring test takers to click on words or phrases.

Instructions: Like MCMR items, multiple-select hotspot items can either specify the number of correct responses or instruct test takers to select all options that apply.

Data considerations: Data from a single-response hotspot item might come in the format of a simple correct/incorrect score, predefined regions specified by content developers, or the XY coordinates that the candidate clicked. For some content, defining regions of interest may be straightforward, such as when an image breaks down into clearly identifiable regions (e.g., countries on a map or regions of the brain). In other instances, such as the MRI in Figure 1, it may be more informative to see where the correct (green) and incorrect (red) clicks were relative to the image.

Figure 1. Correct and incorrect responses identifying brain stem in an MRI



Scoring: Scoring of multiple-select hotspot items should take into account the same considerations as MCMR items.

Fill in the blank

Description: This item type provides a field for candidates to enter their response to the item. Calculation items (fill in the blank items for numerical responses) are considered separately.

Instructions: In our experience over several programs using this item type, test takers provide a wide range of small variations on correct responses. The instructions and format of the item may help reduce the variability of responses, as can direct rather than open-ended questions.

Data considerations: The data returned by this item type is straightforward – consisting of a text string.

Scoring: Our experience has been that the number of correct responses to fill in the blank items is always greater than can be anticipated by item writers, and greater than the number of observed correct answers during pretesting (see Table 2). For this reason we recommend caution in using short answer items unless responses are scored by an advanced algorithm or human raters. Additionally, scores for short answer items should be evaluated based on their probability (given overall ability) so that low-probability incorrect responses can be reviewed. This process helps to identify “unkeyed” correct responses.

Table 2. Number of unique correct and incorrect responses to several items

Item	# unique incorrect responses	# unique correct responses	Total responses
1	26	16	59
2	40	2	56
3	51	2	56
4	35	6	53

Calculation

Description: This item type require test takers to calculate a numeric value and enter that value in a response field. This item type is a subset of the fill in the blank item type, in that only certain characters are allowed (0-9 and decimal typically).

Instructions: For certain types of calculations the precision required should be specified.

Data considerations: Provided calculation items restrict input to numerical values, there should not be any issues with data format. In some cases the units of measurement may be included along with the calculation, which can make scoring and analysis more difficult.

Scoring: When responses may involve various levels of precision, or when there is an acceptable range of responses, calculation items should allow scoring based on numerical ranges rather than strict text comparison. For example, if acceptable responses to an item include values between 2 and 2.5, the scoring should not be based on comparing responses to the values ‘2’, ‘2.0’, ‘2.1’, ‘2.2’, ‘2.3’, ‘2.4’, and ‘2.5’.

Drag and Drop

Description: Drag and drop items involve content that test takers move to respond to a question. For example, an item might involve selecting titles or legends to label a chart, dragging words to complete a paragraph, or dragging images to complete a puzzle.

Data considerations: The analysis of these options is easily accomplished by treating each component as a standard single-response entity.

Scoring: These item types can be constructed with varying levels of dependency, such as when all options have a corresponding match vs. the case when only some options have a

corresponding match. These item types can even be set up with unique sets of options for each component of the answer. Because of the likelihood of local item dependency due to the overlap of options, LID should be evaluated if each component is calibrated/scored as a separate item. Even if the individual parts of an item draw from unique sets of options, a single polytomous score should be considered in place of multiple dichotomous scores because polytomous scoring removes the need to evaluate and remove items based on LID.

Sorting

Description: Sorting items involve categorizing information, such as classifying a list of substances into “ferroelectric” and “ferromagnetic” categories.

Data considerations: From a data format and analysis perspective, sorting items combine the characteristics of multi-select and matching item types.

Scoring: Sorting tasks involving only two categories function as a single multiple-select items, while three or more categories must be analyzed and scored separately.

List Matching

Description: Test takers must connect information between two lists, such as matching states with their capitals or drugs with their classifications.

Data considerations: Data considerations for list matching are identical to those for drag and drop items.

Scoring: List matching items typically have a one-to-one correspondence between lists, and should therefore be scored/calibrated as a single item.

List Ordering

Description: List ordering involves placing information into a sequential list. This might involve ranking the severity of medical conditions, or ordering sentences to create a cohesive story.

Data considerations: Response analysis for list ordering items is somewhat complicated by the fact that meaningful information is not necessarily location dependent. Responses can be analyzed in the aggregate (as with multiple-response multiple-choice items) which yields $L!$ possible combinations. Disaggregating responses by order (e.g., analyzing the element in locations 1, 2, 3, etc.) ignores the relationship between elements. Of particular interest is the analysis of adjacent pairs. Table 3 shows the various ways that the response for a 3 item list ordering item can be broken down for analysis.

Table 3. List Ordering Data Options

	Response					
	123	132	213	231	312	321
Position 1=1	1	1	0	0	0	0
Position 1=2	0	0	1	1	0	0
Position 1=3	0	0	0	0	1	1
Position 2=1	0	0	1	0	1	0
Position 2=2	1	0	0	0	0	1
Position 2=3	0	1	0	1	0	0
Position 3=1	0	0	0	1	0	1
Position 3=2	0	1	0	0	1	0
Position 3=3	1	0	1	0	0	0
Pair1=12	1	0	0	0	0	0
Pair=13	0	1	0	0	0	0
Pair1=21	0	0	1	0	0	0
Pair1=23	0	0	0	1	0	0
Pair1=31	0	0	0	0	1	0
Pair1=32	0	0	0	0	0	1
Pair2=12	1	0	0	0	0	0
Pair=13	0	1	0	0	0	0
Pair2=21	0	0	1	0	0	0
Pair2=23	0	0	0	1	0	0
Pair2=31	0	0	0	0	1	0
Pair2=32	0	0	0	0	0	1

Scoring: For polytomous scoring of these item types, we have found that the using the order of adjacent pairs produces scores with good measurement characteristics. Each element has a correct order associated with it (e.g., the first element has order of 1, the second 2, etc.). Scores are then calculated based on to elements being in the correct order, or the correct relative order (earlier element followed by later element). This produces a score ranging from 0 to L-1, where L is the number of elements in the item. The decision to use absolute or relative order produces very different score distributions, as shown in Table 4. Table 5 Shows the frequency and average theta of actual scores for a list ordering item scored both ways (both scores had comparable item-theta correlations of 0.38).

Table 4. Sample Scoring for 4-Element List Ordering Item

	Relative Order		Exact Order	
Score	Score Frequency	Example	Score Frequency	Example
0	1	4-3-2-1	11	1-3-2-4
1	11	2-4-3-1	9	1-2-4-3
2	11	1-2-4-3	3	2-3-4-1
3	1	1-2-3-4	1	1-2-3-4

Table 5. Score

	Relative		Strict	
Score	N	Avg Theta	N	Avg Theta
0	2	-0.02	32	-0.18
1	27	-0.16	15	-0.06
2	41	-0.09	23	-0.03
3	15	0.04	15	0.04

Item analysis

Standard multiple-choice items are familiar to both item writers and psychometricians, and thus are typically easier to write and analyze. Most item analysis software and syntax has been written with these item types – and these item data – in mind. While the analysis of scores for innovative items is straightforward, working with responses to innovative items often requires advanced data/string manipulation. As new item types are developed, special care should be given to defining the format of response data to allow for easy access to data in the format(s) necessary.

Innovative item types typically present several measurement opportunities from each response. While any item can be scored dichotomously, items requiring multiple decisions or responses may warrant polytomous scoring. Dichotomous items require smaller samples for statistical analysis, have well accepted methods for analysis, are easily understood by stakeholders, and can be easily integrated into test content specifications. Polytomous scoring takes advantage of the additional measurement opportunities provided by complex items, the score points cover a range of candidate abilities, and they are typically more discriminating (higher point biserials) than dichotomously scored items.

Item Response Theory models for scoring polytomous items include the Partial-Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992), and the nominal response model (Bock, 1972). All of these models require greater sample sizes than their dichotomous counterparts. Yen and Fitzpatrick (2008) describe the nominal response model as a method for “analyzing items in which it may not be clear *a priori* which answer choices reflect greater ability” (p117). The nominal response model should therefore be considered a tool that informs the development of item scoring, rather than a method for scoring and equating tests.

The most basic method for determining item scores is through the content expertise of the item writer. In the test development process, there should be exceedingly few instances in which there are no *a priori* correct or incorrect answer choices. This does not mean that the best way of scoring the item is known, or that all of the correct or incorrect answers are known *a priori*. The relationship between a response or the probability of a response and ability (based on an interim score for the test items) provides a basis for evaluating responses.

Figures 2 and 3 show the probability of receiving each score category as a function of ability (theta). The first graph shows a five-point item in which each score category is functioning as expected. As ability increases the probability of the next score point increases, and the

probability of the previous score point decreases, and there is a well-defined threshold where adjacent score points are equally probable. The second figure shows an example where category probabilities are not well ordered – due to problems with one of the score categories. For this item, there is no point at which a four is the most likely score.

Figure 2. Polytomous Item with Ordered Score Probability Curves

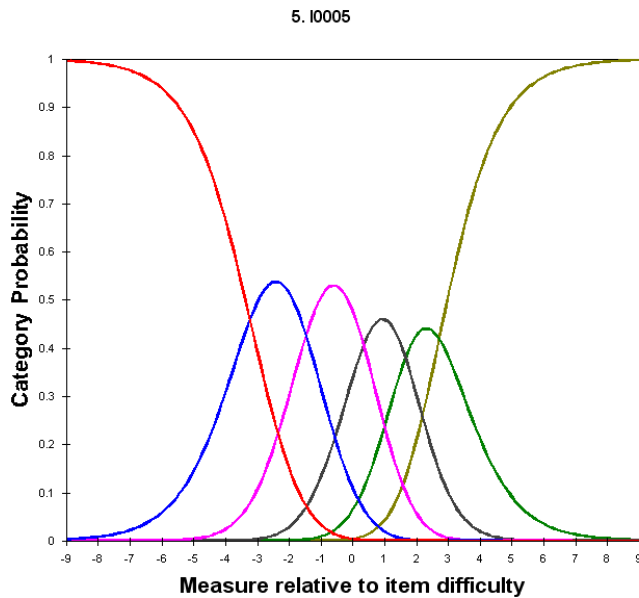


Figure 3. Polytomous Item with Problem Score Category

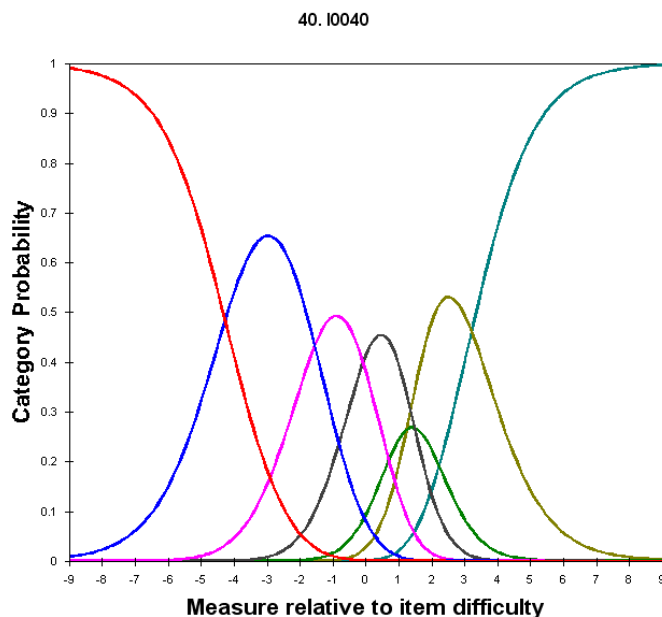


Table 6 shows the descriptive statistics for several scores produced for a set of MCMR items administered with no limitation on option selection (test takers could select as many options as they wished). Dichotomous scoring provides 1 point if all correct options are selected and no

incorrect options selected. The “# Correct” score counts each correct option selected. The “# Incorrect” score is a count of each incorrect option selected. Finally the partial credit score counts correct options selected and subtracts incorrect options selected, with a minimum score of zero. Dichotomous scoring clearly misses measurement opportunities, with all point biserial correlations lower in magnitude than any of the other options. The subtraction of points based on incorrect choices both moderately improves the discrimination of the items relative to correct only and removes the risk of receiving high scores by selecting all options. This type of analysis is particularly useful during the early stages of introducing innovative items.

Table 6. Comparison of Point Biserial Correlations Several Scoring Options

Item	Scoring			
	Dichotomous	# Correct	# Incorrect	Partial Credit
6	0.36	0.52	-0.41	0.59
7	0.41	0.53	-0.46	0.59
8	0.23	0.54	-0.37	0.56
12	-0.04	0.41	-0.2	0.45
13	0.32	0.46	-0.26	0.48
14	0.11	0.48	-0.37	0.54

One advantage of not restricting candidates to a limited number of options is the ability to rescore items. With an explicit instruction on the selection of options for hotspot or multiple-answer multiple-choice items (e.g., “choose 2 options”) the ability to recover scores in the case of miskeyed items is limited. For example, item 12 in table 1 clearly has problems with at least one key (based on the dichotomous point biserial). Because test takers were not explicitly told to select 4 options, that item could be rescored without requiring the collection of new data.

The use of polytomously scored items with an IRT model requires careful evaluation of the scoring categories. Item analysis must include a review of response category frequency to check for unobserved or low-frequency options. Low-frequency score categories lead to inaccurate parameter estimates, while unobserved categories either indicate that the item is scored incorrectly or that it is too easy/difficult for the population. Items with unobserved or low frequency score categories may need to be discarded, or collateral information may be necessary to estimate IRT parameters (Belov & Knezevich, 2008; Hall & Ansley, 2008; Mislevy, Sheehan, & Wingersky, 1993).

While the analysis of responses and score categories on innovative items can indicate if an item is appropriately scored, it is important to ensure that scoring is done in a non-error-prone fashion. For an active testing program there should be a general scoring algorithm for each item type, not individually crafted rules for each item. While the “best” scoring option may vary slightly across items within an item type, the risk that complex scoring brings for correctly building, scoring, and analyzing tests in a large-scale operational setting by far outweighs the potential gains.

There are several common categories of data that may be generated by innovative items. Calculation, fill in the blank, and single-area hotspot items yield single responses which should

not require significant additional processing. Items such as multiple-response multiple-choice, certain hotspots, drag and drop, sorting, ordering and other item types yield multiple responses. While multiple-response item responses can be analyzed as a gestalt, they yield significantly more information if the responses are disaggregated.

Single response items typically function like multiple-choice items with a larger or much larger response space. For multiple response items, the item typically has an overall score (dichotomous or polytomous) based on several related responses (e.g., selecting all true statements, or connecting states with their capitals). While some of these item types could be scored and analyzed as separate items (and sometimes are), it typically makes more sense to score them as a single item due to dependencies between or inseparability of responses. Multiple-response items provide several measurement opportunities, which can be scored polytomously. When scored dichotomously (requiring all correct responses), multiple-response items may be harder than the average single-response multiple-choice item for a program. Items with a polytomous scoring option can also be scored dichotomously based on partial credit (e.g., 1 point for a response that would have yielded 2 or 3 points when scored polytomously). If this is the case, item difficulty can be targeted to key p-value or theta ranges.

Table 7 provides a sample response analysis for a 5 option multiple-select item. Only options/responses with frequency >3 are shown. The disaggregate frequencies show that candidates are selecting all options, which is not immediately clear from the aggregate analysis. However the aggregate data helps to show common misperceptions such as “BCD” which was selected more often than any combination other than the key. For this item, the item-theta correlation for the key (BD) is higher than either of the individual disaggregate options (.20 vs. .09 and .05). In cases where one of the keyed responses is problematic, the item-theta or point-biserial correlation for the fully correct response may be low or negative while many of its components are positive.

Table 7. Option analysis for Aggregate and Disaggregate Responses

Disaggregate				
Option	N	P-value	Avg theta	Item-theta correlation
A	114	19%	0.03	-0.06
B	556	91%	0.20	0.09
C	167	27%	0.03	-0.14
D	573	93%	0.20	0.05
E	118	19%	0.01	-0.14
Aggregate				
Response	N	P-value	Avg theta	Item-theta correlation
BD	317	52%	0.29	0.20
AB	5	1%	0.39	0.04
ABCD	8	1%	0.34	0.03
ABCDE	6	1%	-0.09	-0.06
ABCE	7	1%	0.40	0.05
ABD	39	6%	0.16	-0.03

Response	N	P-value	Avg theta	Item-theta correlation
ABDE	18	3%	0.16	-0.02
ABE	5	1%	-0.33	-0.11
ACDE	6	1%	0.08	-0.03
ACE	5	1%	0.22	0.00
ADE	12	2%	-0.03	-0.07
B	3	0%	0.26	0.01
BC	7	1%	0.00	-0.05
BCD	99	16%	0.09	-0.11
BCDE	4	1%	-0.03	-0.04
BCE	3	0%	-0.40	-0.09
BDE	31	5%	0.18	-0.01
BE	3	0%	0.23	0.00
CD	6	1%	0.39	0.04
CDE	12	2%	-0.05	-0.08
D	9	1%	0.09	-0.03
DE	4	1%	-0.23	-0.08

Table 8 provides an example of possible response analyses for an ordered list item, based on the data considerations discussed in the item type section. Only options/responses with frequency >3 are shown.

Table 8 Option Analysis for Several List Ordering Response Variables

Option	N	P-value	Avg theta	Item-theta correlation
Position1=1	25	29%	-0.42	0.25
Position1=2	44	52%	0.22	-0.23
Position1=3	11	13%	-0.51	0.15
Position1=4	5	6%	0.04	-0.21
Pair1=12	18	21%	0.12	0.27
Pair1=13	4	5%	-0.07	0.05
Pair1=14	3	4%	-0.39	-0.05
Pair1=21	5	6%	-0.12	0.04
Pair1=23	17	20%	-0.36	-0.10
Pair1=24	22	26%	-0.46	-0.20
Pair1=34	9	11%	0.24	0.24
Response=1234	15	18%	0.21	0.31
Response=1243	3	4%	-0.35	-0.03
Response=2134	3	4%	0.29	0.15
Response=2341	16	19%	-0.31	-0.06
Response=2413	5	6%	-0.48	-0.09
Response=2431	17	20%	-0.45	-0.16

Option	N	P-value	Avg theta	Item-theta correlation
Response=3412	6	7%	0.47	0.29
Response=3421	3	4%	-0.22	0.00

Item evaluation

For a program interested in the large-scale implementation of innovative item types, it is critical to understand how different item types perform. Evaluating the comparative quality of innovative item types requires access to accurate classifications of item type in conjunction with the item statistics. A decision may need to be made at the start of the item development process on how item type will be coded – for example, multiple-choice items may contain media stimuli (e.g., audio stems or graphic responses), that are of interest for the evaluation of those items. Especially given the cost of producing media, it is useful to know how these items compare to multiple choice items without a media component. Table 9 provides an example comparison of different item types.

Table 9. Comparison of Item Statistics by Item Type

Item Type	N	Avg Pval	Avg item-theta correlation	% Dropped
Multiple Choice	293	80%	0.290	25%
MC w/media stem or response	38	81%	0.293	29%
Multiple Response	114	78%	0.340	13%
Hotspot	30	68%	0.260	20%
Grand Total	475	79%	0.301	22%

Item statistics reported by item type, while useful, should be interpreted with caution. Especially for new programs, item type can be confounded with item authors, content areas, and usability issues. Additionally, decisions on the use of innovative item types in general, or specific innovative item types, may well depend on factors in addition to average performance. Bearing this in mind, the multiple response items in Table 8 appear to be functioning better than the other item types. In addition to tracking costs, performance of different item types can help plan future item development. With information on item mortality by item type, the number of pretest items necessary to achieve the desired number of approved items can be more accurately estimated.

Because innovative items tend to be more involved than standard multiple-choice items, it is a good idea to keep track of the average time required to complete the item. As programs begin to add new item types to their tests (even as experimental items) they need to realize that these items may take as long as several conventional items. Conversely, if these items are polytomously scored or more discriminating than traditional items, they may provide comparable information to several multiple-choice items. Therefore, in addition to average time per item innovative items should track time per point and time per unit of information as well. Table 10 presents some average item times by item type. The items involving audio stems are longer due

to the length of the recording. Multiple response items are slightly longer than single response items in this item bank.

Table 10. Average Time by Item Type

Item Type	Time
Hotspot w/Graphic Resp	35
MC w/Audio Stem	50
MC w/Graphic Option	34
MC w/Graphic Stem	33
MR w/Audio Stem	57
Multiple Choice	33
Multiple Response	40

Implications for Item Banking

Whether using traditional multiple-choice items (MCQs) or innovative items, an item bank must have the ability to appropriately keep track of relevant item statistics, relevant scoring information must be accessible, and should allow tracking of additional item meta-data. Of particular interest are information regarding media components of items (sound clips, video, audio, and text source), item type, and item author. Keeping track of media components is especially important for identifying items that should not be used together on the same test (unless they are part of a case/testlet). The tracking of item type should also be as molecular as possible – keeping track of stem/prompt media, option media, and response modality. For example, multiple-answer multiple-choice items with graphical stem should be tracked separately from those with text stems.

Like item analysis, most item banking (content management) systems are designed with single-answer multiple-choice items in mind. Probably the most significant difference between traditional multiple-choice items and innovative items with regard to statistics is the response/option analysis. The analysis in Table 8 shows a small portion of the 76 options for which statistics could be generated for a 4 option sorting item. Many of these options will be unobserved (even with large sample sizes), however the unobserved options will vary by item within the item type. Some programs may have more than 4 elements to sort, which will drastically increase the number of potential options (e.g., 5 element list ordering produces 225 possible options).

Implications for Item writing

Because of their higher development costs, innovative items will typically warrant greater revision efforts than traditional item types. The combination of appropriate response analysis and item evaluation can be extremely useful to content developers. Of particular interest is specific feedback on which element(s) of an item with poor performance statistics should be revised to improve performance. In addition to the work required to operationalized the psychometric

analyses, training sessions and materials will be required to familiarize content developers and/or SMEs with the expanded feedback.

Conclusion

Parshall et. al. (2003) point out that innovative items allow us to measure constructs more and/or better than traditional items. The increased number of measurement opportunities, as well as the increased time and expense of innovative items makes polytomous scoring an attractive option. But polytomous scoring has implications for the choice of measurement model, sample size, and test format. Polytomously scored items require greater sample sizes than dichotomous items, which will increase their exposure during pretesting. The increased response space for innovative items reduces the need to estimate a guessing parameter. Finally there is a profound lack of research on item-level adaptive testing using polytomously scored items, and most if not all commercial test administration platforms lack the ability to administer polytomously scored items.

During the initial introduction of innovative items for a testing program concerns such as usability and functionality typically take precedence. While those issues are important, it is equally important to ensure that the data necessary for item level and response level analyses is easily available in a usable format. While response analysis is important for all item types, the increased expense of innovative items makes the revision of items with poor statistics a priority. It is therefore also important to develop new training materials and feedback for content developers and SMEs to ensure that the results of the response analyses are understood and used appropriately.

References

- Belov, D.I., & Knezevich, L. (2008). Predicting item difficulty with semantic similarity measures. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29-51.
- Hall, E., & Ansley, T. (2008). Exploring the Use of Item Bank Information to Improve IRT Item Parameter Estimation. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
- Masters, G. N. (1982) A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, 30(1), 55-78.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Yen, W. M., & Fitzpatrick, A. R. (2008). Item response theory. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Washington, DC: American Council on Education.