Innovative Item Types Require Innovative Analysis

Nathan A. Thompson – Assessment Systems Corporation

Shungwon Ro, Larissa Smith – Prometric

Jo Santos – American Health Information Management Association

Paper Presented at the 2009 Conference of the

National Organization for Competency Assurance

Innovative Item Types Require Innovative Analysis

As computer technology becomes increasingly sophisticated, innovative item types that capitalize on this sophistication are being used in assessments more and more often. The development of such items has received much attention, especially the development of new formats to assess complex constructs with greater fidelity than the ubiquitous multiple choice (MC) item. Examples are as varied as the professions that require assessments for credentialing: medical examinations, computer skills, patient counseling, etc.

However, one important issue that can be overlooked in discussions of "cool" item formats is that data will have to be analyzed psychometrically to ensure the reliability, validity, and defensibility of the assessment. Sound psychometric analysis is necessary to establish the quality of the credential and protect against litigation. Therefore, it is imperative that discussions of innovative item formats include plans for data analysis.

This step can require as much innovation as the development of new formats. Just as item development has historically been dominated by the dichotomously scored multiple-choice item, psychometric analysis has been similarly dominated by models derived to suit those items. Innovative item types, however, require creative forethought and novel methods of applying classical statistics or item response theory.

This presentation will discuss several types of innovative item formats and types of statistics than can be used to evaluate the items and the test. A specific example that will be discussed in depth is a medical coding test where candidates are required to read complex medical cases and determine all applicable medical codes for diagnoses and/or procedures. Candidate performance is evaluated using a complex scoring algorithm requiring an in-depth analysis of each response.

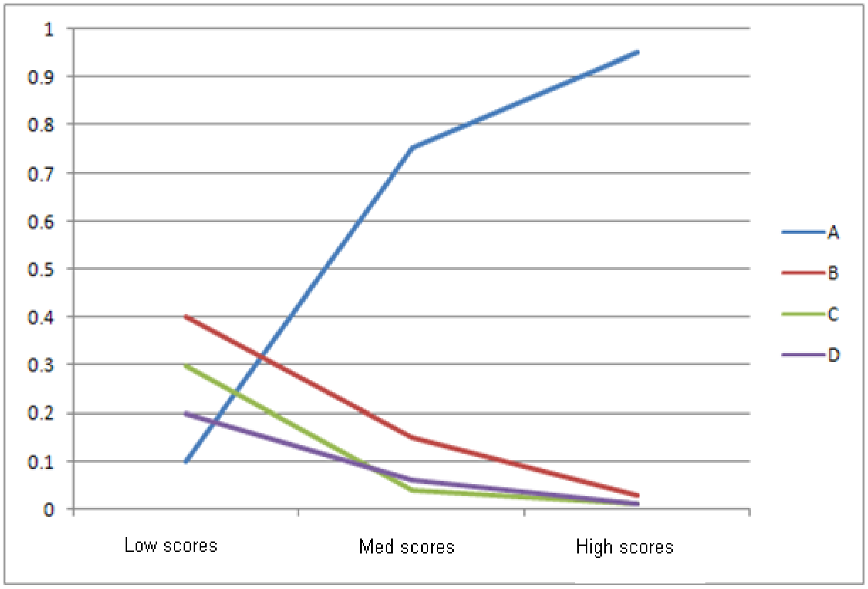
In conclusion, some guidelines will be proposed for the development of innovative item types for credentialing examinations. While the development of formats primarily reflects the intricacies of the profession – which it should, as that is the goal of innovative formats – it must also reflect other considerations like methods of psychometric analysis.

Psychometric Analysis with Conventional MC Items

With classical test theory, multiple choice items are desired to have a positive discrimination (point-biserial or biserial correlation coefficient) for the correct option, with a larger value being better, and the incorrect options desired to have negative values. The proportion endorsing the correct option is often desired to be moderately high, such as 0.40 or higher, with fewer examinees endorsing the incorrect options. It is further desirable that – as a function of a good point-biserial – the proportion endorsing the incorrect option decreases for examinees with higher total scores, while the proportion selecting the correct option increases. This concept of item performance is depicted in Figure 1, where A is the correct response.

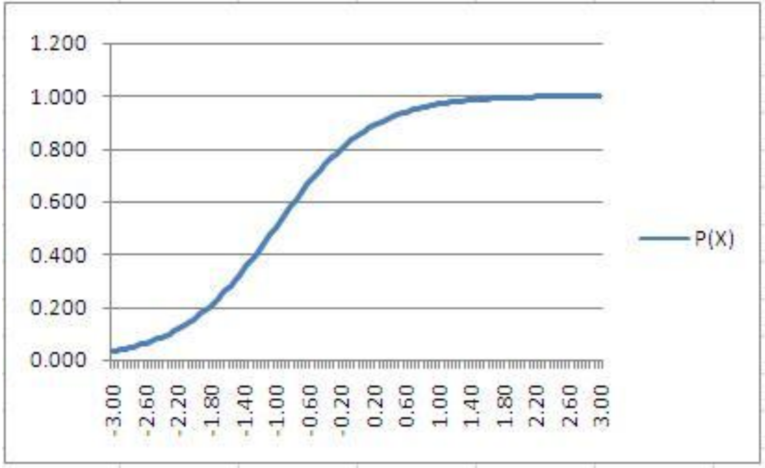
This basic idea of item performance – that examinees with higher total scores or a higher level of underlying ability in the case of item response theory should score better on the item – permeates much of the psychometric analysis of tests. Just as with MC items, the goal of analyzing innovative item data is to show that the items are contributing to quality measurement, which is operationalized in that manner. To this end, an alternative form of item assessment is to find the average score of examinees that endorsed each response; the average score for those selecting the correct response should be the highest in the high score group.

Figure 1: Example Performance Graph of an MC Item



The blue line in Figure 1 forms the theoretical basis of item response theory (IRT), which seeks to find a mathematical model that approximates the probability of correctly responding as a function of ability. But rather than look at three broad categories as Figure 1 does, IRT fits a smooth, continuous line. While the mathematical complexity is much greater, it still attempts to depict the same concept we see in Figure 1 (i.e., the examinee with higher total scores should do better – high probability of answering the item correctly – on the item). An example of an IRT item response function is shown in Figure 2.

Figure 2: Example IRT item response function

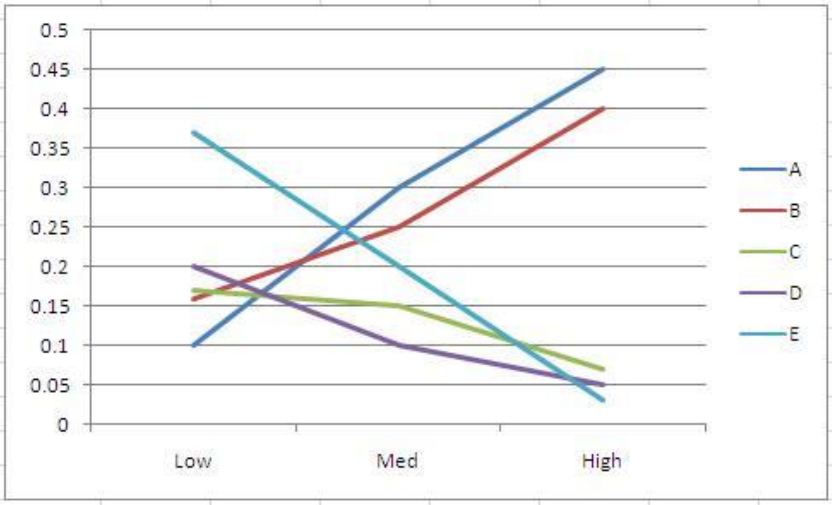


Examples of Alternative Item Formats

Perhaps the first extension of the single-best-answer multiple choice item is the multiple-answer multiple choice item. This type of item is the same as a single-best-answer item except that it requires multiple responses. For example, in medical and allied health professions a case or scenario could be presented, followed by a list of possible patient conditions. The examinee is required to select the two or three most likely or relevant conditions given the scenario.

As the format and development of such items are essentially the same as a multiple choice item, the psychometric analysis follows suit. Correct options are still desired to have positive point-biserials whereas incorrect options have negative point biserials, and that not too many examinees are drawn to the incorrect options. Figure 2 is an item performance graph for this type of item with A and B as the correct answers.

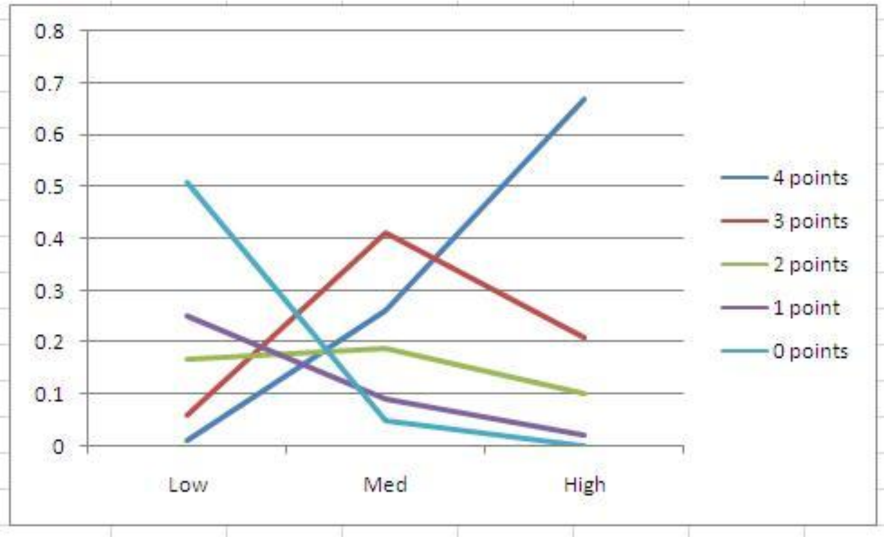
Figure 3: Item performance graph for a hypothetical multiple-response item



Another simple extension is the partial credit item. Examinees are asked to solve a problem, and awarded points by how far they proceed correctly, how cogent the solution is, and in the case of educational tests, whether they show all their work on paper. In some cases these might be scored objectively, but in many cases they are scored subjectively with scoring rubrics. Either way, examinees often receive one of a sequential number of possible points, such as 0, 1, 2, 3, and 4.

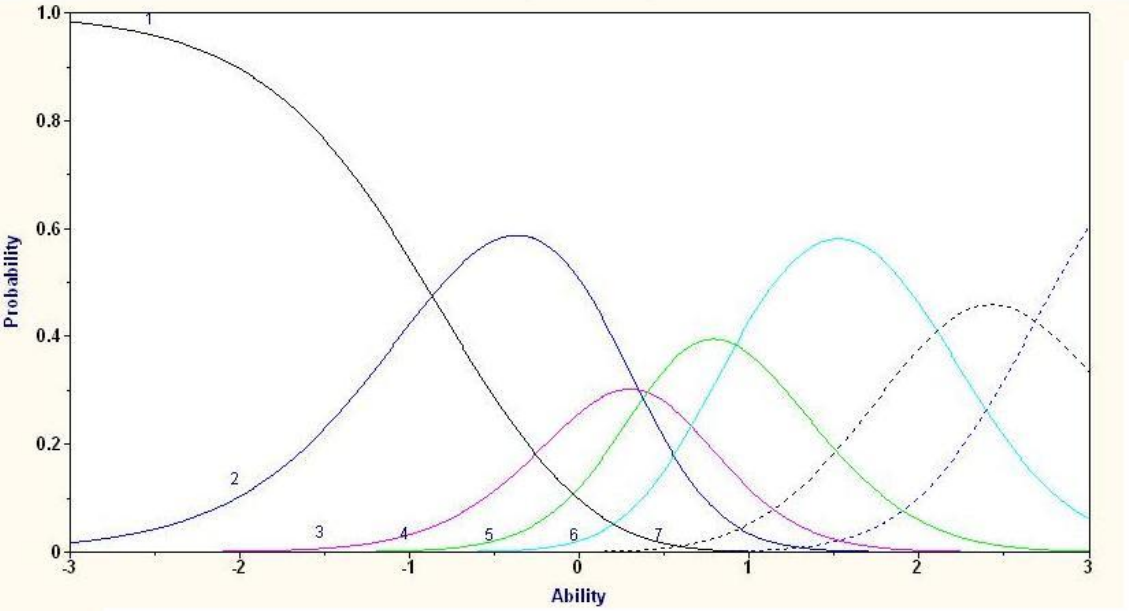
With such an item, we still hope to find that examinees of higher ability score better on the item. A simple method to examine this is to calculate the average total score for examinees in each category. We hope to see this average increase as a function of item score, so examinees with the highest possible points on the item obtaining the highest average total score. This in turn implies a positive polyserial correlation, which is the polytomous (multi-point) version of the biserial correlation. Again, we want to find the proportion in the correct (in this case, highest-point) category increasing as a function of total score. An example of this is presented below.

Figure 4: Example performance graph for a partial credit item



There also exists a branch of IRT devoted to the analysis of polytomous items, which includes rating scale type items as well as partial credit type items. Like the dichotomous (correct/incorrect) function shown in Figure 2 is an extension of Figure 1, the polytomous IRT graph in Figure 5 is a continuous extension of Figure 4.

Figure 5: Example polytomous IRT function for a partial credit item



Many innovative item types can be categorized as a partial credit item, because no matter their format, many are score by awarding a sequential number of points. This assumes that we can take a complex item and reduce it to a simple number of steps, which can be seen either as an advantage or a disadvantage.

Complex Items

Some innovative test items assume that the construct they are measuring is too complex to face a possibly oversimplified reduction to a small number of discrete points. For such items, idiosyncratic point schemes are typically devised. These systems can specifically take into account the order of steps needed to complete an item. Or they look only at whether the steps were completed successfully. They can also be designed as increasing or decreasing systems, so that a score can start at zero and points are added for correct steps, or alternatively that a score starts at a large number like 100 and points are subtracted for error. In either case, the additions or subtractions are commensurate with the importance of the relevant step.

This leads to scores that are able to make a much finer distinctions among examinees. Take for example a four-step complex item that instead of simply being scored 0-1-2-3-4 is scored with the following weight point system:

Step 1:	5 points
Step 2:	8 points
Step 3:	10 points
Step 4:	15 points

The possible scores from this system are 0, 5, 13, 23, 38 if only considered in succession like the 0-1-2-3-4 approach. But with the weighted system other scores are available, like 8, 10, 15, 18, 20, 23, 25, 28, 30, and 33.

This large number of possible item responses makes it difficult to evaluate the item. The simplest statistics to examine would be the number or proportion of examinees obtaining each score. A graph such as Figure 4 could be constructed, but it would be difficult to read because there are now $2^4 = 16$ possible scores, and therefore 16 lines superimposed on one another. An alternative to this would be to plot the average total test score for the examinees achieving each of the possible item scores, which should be an increasing curve. However, this faces the disadvantage that with so many possible item scores, there are likely to be a few item scores that have few or no examinees and therefore not enough data to provide accurate information for evaluation.

Another option would be to treat each step as an individual dichotomous item for the purposes of analysis, ignoring the dependency among the steps. A score of 0 or 1 for each step could be correlated to both the item total score and the test total score. Like a true dichotomous item, a negative point-biserial would indicate a possible issue with the step because high-scoring examinees are doing worse on the step than low-scoring examinees. Likewise, the proportion of examinees successfully completing each individual step should be examined

The fact that there are already 16 possible scores for this item even though it is only composed of four discrete steps presents somewhat of an obstacle for analysis simply due to the sheer number of possible things to examine. Items that are extremely complex or open ended can lead to even more possible item scores, and a very wide range of possible test scores even on a test with only a few items. The following section provides an actual example of a test with a complex scoring system in this model, and some of the statistics that are examined during the test development cycle.

An Example of a Complex Item: Medical Coding

The American Health Information Managers Association (AHIMA) has a number of examinations intended to certify medical coders in various contexts. Medical records coders assign codes to the diagnoses and procedures listed in patient medical records. These codes are then used to determine, among other things, the amount paid to the health care provider by the insurance company. Incorrect codes can lead to errors in reimbursement or even result in patients being denied reimbursement for care; therefore, it is critical that medical coders not only provide correct codes but refrain from recording inaccurate diagnoses or billing for procedures that were not actually performed.

AHIMA has two certification exams with novel item types: the Certified Coding Specialist (CCS) exam and the Certified Coding Specialist – Physician Based (CCS-P) exam. Both exams present innovative test items associated with case study scenarios and require candidates to code those case study items as if they were medical records presented on the job. Responses are open-ended and require candidates to enter codes into boxes on the screen. To maximize the validity of the exam as it relates to the job, candidates are penalized both for providing incorrect codes and for failing to provide correct codes.

The scoring structure of the CCS and CCS-P exams reflects the contents of the task. Candidates are awarded three points for each correct diagnostic and procedure code. For each incorrect code supplied, and for each keyed code not presented in the item response, candidates are penalized three points. In the CCS, which assesses a task in which a primary diagnosis must be supplied, candidates are awarded six bonus points for giving the correct primary diagnosis but are not penalized for failing to provide it. One consequence of this scoring structure is that the range of points possible per item is variable and depends not only on the number of keyed responses but on the amount of physical space in the response area: candidates can only give as many correct answers as are listed on the key, but they can give as many incorrect answers as they have room to give, at a three-point penalty for each incorrect answer.

Take as an example a hypothetical item that has four correct diagnostic codes and two correct procedure codes, administered in a physical space such that there are ten answer spaces for the diagnostic codes and five for the procedure codes, with no bonus given for the correct primary diagnosis. A candidate who gives all correct responses and no wrong responses would receive (4+2)*3=18 points. A candidate who gave none of the keyed responses and as many incorrect responses as the space will allow would receive (4+2)*-3 = -18 points for failing to put the correct responses, plus (10+5)*-3 = -45 points for putting incorrect responses in all the available answer slots, for a total item score of -45-18= -63 points. The item in question therefore has a possible points range of -63 to 18, in increments of three.

A different item with five correct diagnostic codes and three correct procedure codes would have a possible lower bound of [(5+3)*-3] + [(10+5)*-3] = -69 points and a possible upper bound of (5+3)*3 = 24 points. Because items are scored in increments of three, the first item has 28 possible score points and the second item has 32. The sample size for these exams is fairly large over the administration window, but the data set rarely contains all possible score points, in large-enough quantity, for every item.

The proportion of correct and incorrect codes can be calculated by code, as the code-to-total correlations (point biserials) for each code can be evaluated. It is not impossible to analyze this type of item using polytomous IRT models, dealing with very large number of codes is challenging.

An Example of a Complex Item: Situational Judgment Items

A certification exam for a financial institute consists of a section of 80 multiple choice items and a section of situational judgment items. The situational judgment section contains four scenarios, and each scenario includes three questions in select-all-that apply format, with ten options and between three to six correct responses. The situational judgment scenarios are scored according to the number of correct and incorrect options selected, and therefore partial credit is recognized.

For each situational judgment scenarios, ten points are assigned. So with three questions in a scenario, the final score for each response pattern to each question is weighted to 1/3 of the maximum points of 10. As an example, consider a question that has six options that should be answered as ABCDEF and four options left blank (GHIJ). Each correct option is worth of 10 points if the option is selected or 0 point if not selected. Each incorrect option is worth of 1 point if the option is selected or 0 point if not selected. Thus, a candidate who selects 6 correct options and 4 blanks would get 60 points, a candidate who selects 5 correct and 1 incorrect options would get 51 points, a candidate who gives 4 correct and 2 incorrect options would get 42 points, etc. Each candidate's score is then scaled to 10 for a raw score of 60, 8 points for a raw score of 51, 6 points for a raw score of 42, etc. These scale scores for each question are then multiplied by 1/3 to give a maximum of 10 points assigned to each situational judgment scenario. The maximum total score of four situational judgment scenarios will be 40. The total exam score (120) is the sum of multiple choice item scores (80) and the situational judgment scenario scores (40).

Developing Innovative Items Reflecting Innovative Analysis

There have been efforts in the field to go beyond MC item types for credentialing exams. One example is to use the scenario based items to increase the fidelity of the stimulus and increase the number of options to discourage guessing (Clauser and Case, 2006). The use of scenario based or case study items should be considered in the context of appropriate psychometric analyses and scoring. Another example is to use simulations delivered on computers and internet. Technology allows us to reduce the length of the simulation for collecting more cases or more content. Careful selection of scoring method would support this type of innovative items in the computer delivery mode.

Oral examinations have been prevalent in the medical exam settings. It usually requires good sampling of situations and raters necessary for scoring the objects/simulations/scenarios in order to achieve a reasonable level of reliability. With computer delivery of stimulus, a simple or complex scoring model can be utilized to analyze the items and candidates' performance. But scores may remain questionable if sampling of the content and the raters are poor in quality, although more in depth analysis to tab into examinees' thought process is feasible.

Short answer, extended answer and essay exams are often considered appropriate ways of measuring skills and abilities in the credentialing examinations. Research studies often raise concerns on these formats unless the scores on these types of items are combined with scores obtained from other assessments for making high stake decisions. Using MC items in addition to these types usually enhances the reliability and possibility of equitable scores across administrations.

Automated scoring on discrete tasks has been gaining popularity. Depending on the complexity or types of innovation, tasks can be dichotomously scored or polytomously scored with partial credit. Scoring rules to define correct solutions to tasks can be drawn from the experts as seen in the example above. The rules-based approach can be used for automated scoring on more complex item types that require human raters. Research studies show automated scoring algorithms generate fewer errors than human raters in some cases. For economical reasons, some types of innovative items require automated scoring over human scoring. Design and development of those item types should be considering supportive of innovative scoring method like automated scoring methods with rules identified by experts.

There are distinguished benefits of developing and incorporating innovative items into a testing program. That includes, expanded measurement of the construct (skills range) by using visuals and audio and computer technologies as examples, presentation of complex construct, reducing construct-irrelevant variance (for example reducing reading by presenting graphics and animations), reduced guessing, assessing in-depth level of skills using technology based simulation, etc.

However, the process of developing and introducing innovative items should be thoughtful and purposeful. Poor user interfaces and unclear response action requirements can cause poor content validity by construct-irrelevant variances. It is also important to note that much less psychometric information is available for the new item types being developed. Parshall and Harmes (2009) proposed four activities that are useful for the design and development of innovative item types. They are template design, item writing guideline for each new type of items, item writer training and usability check. Like many researchers, they provided practical guidelines for the development of innovative item types, but the aspects of psychometric analysis and scoring were not very much emphasized.

For effective incorporation of innovative item types into testing program, the development process of those items must consider appropriate psychometric analysis. The followings are suggested steps for consideration in this purpose (Parshall, et al, 2000).

1. Clear definitions of a target construct – Specify the purpose of the assessment and define the construct that is expandable with advanced technology and innovative item types and analysis.
2. Skills and tasks defined – specification on individual tasks for introducing innovative item types.
3. Item formats designed – specify innovative item format other than traditional MC (single or multiple response, ordered response and matching items) including short and extended answer, essay, and simulation targeted features/responses.
4. Response action defined – consider and specify the exact level of constraints in requirement of examinees action to respond to items (reaction).
5. Inclusion of media other than text – consider using media such as graphics, audio, video and animation to improve task congruence and validity of the constructed exam and reducing content-irrelevant variance by limiting reading requirement of texts.
6. Determination of interaction level between stimulus (for example, computer simulation) and the examinee – appropriate levels of interaction needs to be considered for feasible scoring and analysis.
7. Select appropriate scoring algorithm and model – consider simple automated scoring by matching keyed responses to complex automated scoring with rules defined and weights determined by the experts.

Advanced technologies make more complex item types feasible. Innovative item types can be often challenge for psychometric analyses unless good features of such items can be artificially simplified at times. Without considerations on appropriate psychometric methods to use, innovative item types may not well be serve the purpose of assessment with necessary evidences of validity and reliability.

References

Clauser, B. E., Margolis, M. J., & Case, M. C. (2006). Testing for Licensure and Certification in the Professions. *In Educational Measurement, 4th Ed. Edited by Brannan, R.* American Council on Education and Praeger Publishers.

Mills, C. N., Potenza, M. T., Fremer, J. J., Ward, W. C. (2002). *Computer –Based Testing: Building the Foundation for Future Assessments*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Downing, S. M., & Haladyna, T. M (2006). *Handbook of Test Development*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Parshall, C. G., & Harmes, J. C. (2009). Improving the Quality of Innovative Item Types: Four Tasks for Design and Development. *Journal of Applied Testing Technology*, Vol. 10, 1.

Parshall et al. (2000). Innovative Item Types for Computerized Testing. *In Computerized Adaptive Testing: Theory and Practice. Edited by van der Linden, W. J., & Glas, C. A. W.* Kluwer Academic Publishers, Dordrecht/Boston/London