Yardstick
Testing and Training Experts

Scoring Models for Innovative Item Types: A Case Study

By Dr. Chris Beauchamp **January 2016**

As a way to capture the richness of job performance, many credentialing organizations are supplementing traditional multiple choice questions (MCQs) with innovative item types. Although this view is not unanimous, the theory is that MCQs represent a somewhat artificial representation of job tasks, and that these innovative items represent a more refined way to assess candidate competence.

There are a number of benefits and challenges related to innovative item types. One of the biggest challenges is related the selection of an item scoring methodology. The purpose of this backgrounder is to discuss a recent project where four different scoring models were used to score candidate response data to a pilot project that includes four item types.

The Pilot Test

This project involved the administration of a 70-item test to a group of 71 volunteers. The test involved four item types:

- Traditional Multiple Choice
- Multiple True-False
- Drag and Drop to Classify
- Drag and Drop to Order

Traditional Multiple Choice (47 items)

The candidate is asked to select the single correct or best answer from the 4 options provided. There are 4 answer options for every item.

Example: Who was Canada's first Prime Minister?

1. Sir John A. McDonald (*)
2. William Lyon Mackenzie
3. George Washington
4. Mackenzie King

Yardstick
Testing and Training Experts

Multiple True-False (9 items)
The candidate is presented with a stem and 4-9 answer options. For each option, the candidate is asked to indicate if the statement is true or false.

Example: Which of the following teams were part of the "Original Six" teams in the National Hockey League (NHL)?
1. The California Golden Seals (False)
2. The Philadelphia Flyers (False)
3. The Boston Bruins (True)
4. The New York Rangers (True)
5. The New York Islanders (False)
6. The Detroit Red Wings (True)
7. The Colorado Rockies (False)

Drag and Drop to Classify (10 items)
The candidate is presented with instructions followed by a list of 4-7 options. For each option, the candidate is asked to indicate which category applies to the option.

Example: Based on the Canadian Constitution, identify if the following fall under federal or provincial jurisdiction:
1. Healthcare (provincial)
2. Military (federal)
3. Education (provincial)
4. Telecommunications (federal)
5. Criminal Code (provincial)

Drag and Drop to Order (4 items)
The candidate is presented with instructions, followed by a list of 4 options. The candidate is asked to place the options in order.

Example: Place the following dance movies in the order in which they were released:
1. Top Hat (1)
2. Step Up (3)
3. Dirty Dancing (2)
4. Bring it On (4)

Yardstick
Testing and Training Experts

The Scoring Models

Candidate responses were scored using four different scoring models:

- Dichotomous Model
- Modified Dichotomous Model
- Partial Credit Model
- Negative Scoring Model

Dichotomous Model

Candidates receive 1 point if they answered every element of a question correctly. They receive 0 points if they answer any element of the question incorrectly. As a result, candidates receive a score of 1 or 0 on each item.

Example:

Stem: Based on the Canadian Constitution, identify if the following options fall under federal or provincial jurisdiction:

Option	Candidate Response	Correct Response	
Healthcare	Provincial	Provincial	✓
Military	Federal	Federal	✓
Education	Provincial	Provincial	✓
Telecommunications	Federal	Federal	✓
Criminal Code	Provincial	Federal	✗

Candidate Score: **0 points**

Yardstick
Testing and Training Experts

Modified Dichotomous Model

As a variation to the dichotomous model, candidates receive a point if they achieved a score of 75% or more on an item. They receive 0 points if they achieved a score of less than 75%. As a result, candidates either receive a score of 1 or 0 on an item.

Example:

Stem: Based on the Canadian Constitution, identify if the following options fall under federal or provincial jurisdiction:

Option	Candidate Response	Correct Response
Healthcare	Provincial	Provincial ✓
Military	Federal	Federal ✓
Education	Provincial	Provincial ✓
Telecommunications	Federal	Federal ✓
Criminal Code	Provincial	Federal ✗

Candidate Score: **1 point (candidate achieved a score of >75% on the item)**

Yardstick
Testing and Training Experts

Partial Credit Model

Candidates receive partial credit for each option they answer correctly. They receive a score of 0 for each option they answer incorrectly. As a result, candidates can receive a score anywhere between 0 and 1 based on how well they performed on the item.

Example:

Stem: Based on the Canadian Constitution, identify if the following options fall under federal or provincial jurisdiction:

Option	Candidate Response	Correct Response	
Healthcare	Provincial	Provincial	✓
Military	Federal	Federal	✓
Education	Provincial	Provincial	✓
Telecommunications	Federal	Federal	✓
Criminal Code	Provincial	Federal	✗

Candidate Score: **0.80 points**

Yardstick
Testing and Training Experts

Negative Scoring Model

Candidates receive partial credit for each option they answer correctly. They will receive negative credit for each option they answer incorrectly. As a scoring rule, with the exception of MCQ, candidates cannot receive a score of less than 0 on an item. As a result, candidates can receive a score anywhere between 0 and 1 based on how they responded to the item.

Example:

Stem: Based on the Canadian Constitution, identify if the following options fall under federal or provincial jurisdiction:

Option	Candidate Response	Correct Response	
Healthcare	Provincial	Provincial	✓
Military	Federal	Federal	✓
Education	Provincial	Provincial	✓
Telecommunications	Federal	Federal	✓
Criminal Code	Provincial	Federal	✗

Candidate Score: **0.80 points - 0.20 points = 0.60 points**

Yardstick
Testing and Training Experts

Results

Candidate scores and test-level metrics were calculated based on the application of the four scoring models to the test.

Example:

	Dichomotous	Modified Dichotomous	Partial Credit	Negative Scoring
Number of candidates	71	71	71	71
Number of items	70	70	70	70
Average score	38.15	42.89	46.65	38.89
Standard deviation	7.16	7.61	7.49	7.68
Minimum score	11.00	13.00	13.23	10.00
Maximum score	50.00	55.00	56.74	51.48
Average item difficulty	0.55	0.61	0.67	0.56
Average discrimination	0.22	0.24	0.29	0.20
Reliability	0.81	0.82	0.86	0.79

Candidate scores

Candidates achieved higher scores using the partial credit model. Scores were lowest for the dichotomous and negative scoring models. Based on the intended use of this test, one could argue that the dichotomous and negative scoring models led to unexpectedly low scores that may not be a true reflection of a candidate's ability.

Test Reliability

Reliability for the test was calculated using Cronbach's alpha coefficient. This measure produces an index that ranges from 0 (no reliability) to 1 (perfect reliability). In credentialing examinations, the minimum acceptable reliability is above 0.80 (Nunnally, 1978). The partial credit model produced the best reliability (0.86). The other three models yielded lower reliability, with the worst model being negative scoring (0.79).

Yardstick
Testing and Training Experts

Item-level performance

In addition to item difficulty (i.e., the percentage of candidates answering a question correctly), one important metric to assess item performance is discrimination. Discrimination refers to an item's ability to differentiate between those who know the content and those who do not. Items with higher discrimination do a good job of identifying competent candidates versus not-yet-competent candidates. For this study, the corrected point biserial correlation (CRPB) was used. This measure produces an index that ranges from -1.00 (perfect negative discrimination) to 1.00 (perfect positive discrimination) with an index of 0.00 indicating no discrimination. Based on average discrimination, the partial credit model was most effective in differentiating between competent and not-yet-competent candidates. The negative scoring model yielded the least discriminating items.

Conclusion

In many ways, the partial credit model is the simplest and most intuitive model. It yielded scores that best match the true representation of a candidate's ability, displayed evidence of higher test reliability, and produced the best average indices of item discrimination. In addition, the partial credit model, unlike the negative scoring model, does not penalize candidates for guessing. Within the context of this particular test and this particular candidate population, the evidence supports the use of the partial credit model over the other three models.

References

Nunnally, J.C. (1978). Psychometric Theory, 2nd ed. New York: McGraw-Hill.

To read more articles related to eLearning, examination, and instructional design go to www.getyardstick.com and check out our blog.