# Evaluating Scoring Rubrics for Innovative Item Types

**Ally Thomas**
*The Graduate Center*
*CUNY*
astevens@gc.cuny.edu

**Jay Verkuilen**
*The Graduate Center*
*CUNY*
jverkuilen@gc.cuny.edu

**Howard Everson**
*The Graduate Center*
*CUNY*
heverson@gc.cuny.edu

The emergence of innovative item types for the next generation of large scale assessments brings new measurement challenges and issues. In the summer of 2013, technical reviews were released regarding the relative progress of the PARCC and Smarter Balanced consortia in developing assessments aligned to the Common Core State Standards. Reviewers strongly suggested that further research be conducted on the innovative items, including developing evidence of their validity with a focus on the utility and validity of their scoring rubrics. Our study attempts to address this validity challenge by investigating (1) students' typical response patterns (2) whether their corresponding scoring rubrics validly measure the students' responses to these item types.

One particular innovative item type of interest is the multiple selection item. This item type provides students with several options as potential answers, and asks them to "select all that apply". It may appear that multiple selection item types are hardly innovative—nothing more than an extension of multiple choice. But, this seemingly simple extension incorporates new complexities. The option of there being multiple answers instead of the typical one right answer, requires a higher order level of skills to correctly answer the item (Ackerman, Evans, Park, Tamassia & Turner, 1999; Mills, 2000). In addition, the number of options examinees must decipher through adds to the cognitive complexity (e.g., the greater the number of response options the greater the cognitive load). These facets subsequently lead to a high variability of response patterns and an increase in the complexity of how the item should be scored.

To facilitate the discussion the below exemplar item (figure 1), "the field trip" item, will be used to illustrate the potential scoring issues that may arise. The field trip item asked students to identify the correct combination of vehicles needed to take three classes on a field trip. Five options were given, and students were asked to "select all that apply". The correct answer included selecting options A, C & D and points were allocated based on the following scoring rubric:

- 0 if no correct answer choices were made out of any possible number submitted
- 0 if all five choices were submitted
- 0 if two correct and two incorrect choices were made out of four submitted
- 1 if three correct answer choices were made out of four submitted
- 1 if two correct answer choices were made out of three submitted
- 1 if two correct answer choices were made out of two submitted
- 2 if three answer choices made were correct out of three submitted

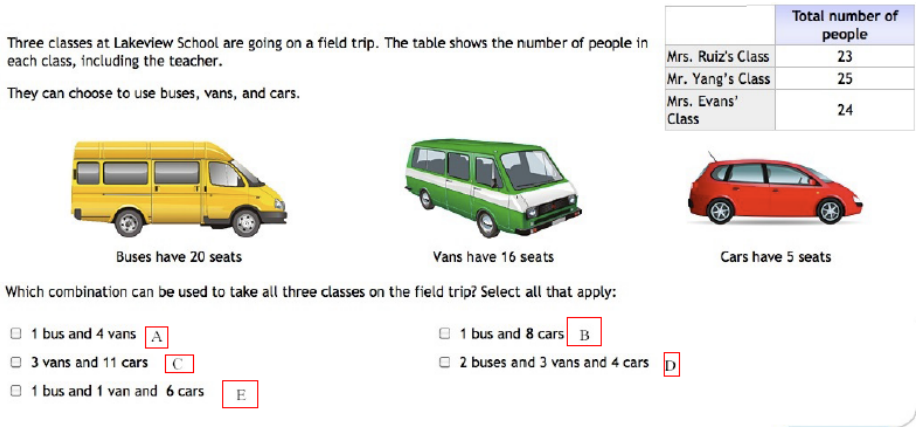


*Figure 1.* Exemplar item: the field trip.

The field trip item was administered to 243 fourth grade students in during a pilot study of innovative mathematics items. The majority of students identified as Hispanic (56.4%) with 29.6% White, 4.9% Black, 4.5% American Indian or Alaskan Native, 1.2% Asian or Pacific Islander, and 3.3% multi-racial. Furthermore, state assessment scores revealed that 65.2% were proficient or above in reading and 67.1% were proficient or above in math.

Latent Class Analyses were conducted in an effort to classify and assess students' response patterns which was then compared to the scoring rubric. The field trip item revealed a four class solution. *Class 1* had a high probability of selecting option E (74%), *class 2* had a high probability of selecting options A (49%), C (100%) and D (60%), *class 3* had a high probability of selecting option A (100%), and *class 4* had a high probability of selecting option D (100%). After investigating these analyses we now understand further students' responses to the item; students tended to select either the correct answer (A, C, & D) or only one option (A, C, D or E).

Examining students' response patterns gives evidence that the underlying assumptions of the scoring rubric were incorrect. A large number of students selected only one option, giving them zero points regardless of whether that option was one of the three possible correct options (e.g., the correct options were A, C & D and the student selected only option A, but received zero points). The scoring rubric did not account for students responding in this manner and as a result only 11% of respondents received full credit (two points), 15% received partial credit (one point) and 74% received no credit.

Adapting the original rubric to allow partial credit for the selection of one single correct answer and no other options selected, dramatically decreased the percentage of respondents earning zero points (table 2). Now only 26% of respondents earned zero points as compared to 74% under the original rubric. The original scoring rubric made it appear that students' performance on this item was abysmal, but in reality nearly half (48%) of the students were able to identify at least one correct option.

**Table 2**
**Comparison of Latent Class Membership and Points Earned in Original and Adapted Rubrics.**

| | **Points Earned** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Original Rubric | | | Adapted Rubric | | | |
| **Latent Class** | **0** | **1** | **2** | **0** | **1** | **2** | **3** |
| E (*n* = 47) | 47 | 0 | 0 | 47 | 0 | 0 | 0 |
| A, C & D (*n* = 71) | 23 | 22 | 26 | 1 | 22 | 22 | 26 |
| A (*n* = 43) | 43 | 0 | 0 | 2 | 41 | 0 | 0 |
| D (*n* = 70) | 58 | 12 | 0 | 9 | 49 | 12 | 0 |
| TOTAL | 171 | 34 | 26 | 59 | 112 | 34 | 26 |

This example makes explicit the importance of incorporating supporting evidence into the design of innovative complex item types. The original scoring rubric created for the field trip item inferred that almost no students had the knowledge, skills or abilities to answer the item correctly. But, after examining further the students' responses it is clear that the scoring rubric lead to invalid inferences—many students had the knowledge, skills and abilities to identify at least one correct response. In addition, incorporating partial credit scoring in instances such as these may increase the information yield per item and subsequently improve the items' discriminatory ability. Furthermore, these analyses suggest that students may not understand the instructional set for these novel items, and their novelty may lead to less than valid measurement. If item types like these are to be used in the forthcoming next generation assessments, there will be a need for further investigation of these potential validity threats, as well as a need for a more systematic approach to studying the validity of scoring rubrics.