Journal of Applied Testing Technology, Vol 22(1), 37-51, 2021

Psychometric Properties of Alternative Item Types Worth the Squeeze? An Investigation into the Psychometric Performance of Alternative Item Types

Amanda A. Wolkowitz[1*], Brett P. Foley[2] and Jared Zurn[3]

[1]Senior Psychometrician, Alpine Testing Solutions, Inc., 51 W. Center Street, #514, Orem, UT 84057, United States; Amanda.Wolkowitz@alpinetesting.com

[2]Director of Professional Credentialing and Senior Psychometrician, Alpine Testing Solutions, Inc., 51 W. Center Street, #514, Orem, UT 84057, United States

[3]Vice President, Examination, National Council of Architectural Registration Boards, 1401 H St NW #500, Washington, DC 20005, United States

Abstract

As assessments move from traditional paper-pencil administration to computer-based administration, many testing programs are incorporating alternative item types (AITs) into assessments with the goals of measuring higher-order thinking, offering insight into problem-solving, and representing authentic real-world tasks. This paper explores multiple applications of AIT items and the psychometric properties of these items, including item response time, difficulty, item-total score correlations, and distractor analyses. The appropriate use of these items is also discussed in the context of professional credentialing exams.

Keywords: Alternative Item Types, Innovative Item Types, Item Development, Item Analysis and Credentialing

1. Introduction

Nearly 25% of all working Americans hold at least one certification or license (Bureau of Labor Statistics, 2018). These credentials span a wide range of industries and job roles, and are awarded by many credential-granting organizations. These organizations monitor their respective industries to make sure their credentialing exams reflect current job expectations. Moreover, many of these organizations are considering innovative item types for their exams to assess relevant content in new ways. Innovative item types, also referred to as Alternative Item Types (AIT), typically describe items that are not traditional standalone, multiple-choice items.

Hotspot (HS) items provide a good example of how a traditional, multiple-choice item can evolve into a more efficient item through technology. Hotspot (HS) items present a question followed by a drawing, photograph, diagram, or other image. To respond to this item type, candidates click on an area or object within the provided image. If they click on the correct area of the graphic, they earn the point for the item. In a multiple-choice item, these items would have pre-defined regions (e.g., A, B, C, D) and the choices would be each region. As a technology-enhanced HS item, the discrete options are eliminated and candidates are faced with a question more aligned to a real-life situation, e.g., they must select the correct region in a graphic without being presented with a list of possible responses.

The primary purpose of a HS item or any exam item is to measure a candidate's knowledge in the construct of interest. As described in the example above, AIT items use technology to enhance the measurement opportunity (Sireci & Zenisky, 2006). It is important to recognize that innovation is not the goal in AIT; instead, innovation is the means to an end. In the hotspot example above, the use of the AIT enhanced the opportunity to measure a candidates' knowledge of a situation. If it failed to increase

*Author for correspondence

the measurement opportunity, then the use of that item type would not be the most appropriate way to assess the content. As such, the efficacy and efficiency of AIT items should be assessed and validated (Haladyna & Rodriguez, 2013).

Parshall and Harmes (2008) discuss a six-step process for determining whether to use AIT items. They suggest that the first step is to determine if there is a need. Test developers and subject matter experts should carefully review the test specifications and determine if any areas cannot be successfully assessed with traditional, multiple-choice item types. If so, then test developers and subject matter experts should determine which item types should be used to fill that need. The next step is to develop prototypes for that AIT and continue to review and refine the prototype until it fits the area of need.

Once the prototype is molded to fill the identified need, the next step is to develop pilot items (which requires documents such as item writing manuals, item writer training presentations, and templates) and beta test the AIT items. An item analysis should be conducted after the beta test to determine if the items are performing in an acceptable and desired way. If not, the AIT items should be reconfigured (with updated documentation) and re-piloted. After the beta test, the final step in the process is to document any final decisions and begin implementing the new AIT items into the exam.

2. Types of AIT Items

Items tend to classify into one of three categories: selected-response, constructed-response, or performance-based responses. Table 1 summarizes the types of items in these categories and when they are appropriate to use. In general, most AIT items tend to fall into either the constructed-response or performance-based item response format. However, an item may be considered an AIT item if the stimulus has been enhanced with technology to make it more real-life. For example, an exam covering architecture may have a technology-enhanced scenario that provides multiple pieces of information about an architectural project (e.g., client request, building codes, site plan, 3-D model of prospective building). In this case, the scenario is the enhanced portion of the item. Although an AIT item may certainly accompany a scenario such as this one, a traditional, multiple-choice item may also serve the purpose. Again, AIT items or stimuli should be used to enhance items for the purpose of improving assessment and not just adding variety to the exam.

Table 1. Item Type Categories

Item Format	Item Types	Appropriate Use
Selected-Response	Multiple-Choice Select-All-That-Apply True/False	To measure knowledge, skills, and abilities across a broad content domain. To measure high-level thinking skills, such as problem solving, analyzing, synthesizing, or evaluating.
Constructed-Response	Short answer (e.g., quantitative-fill-in-the-blank) Essay	To measure knowledge, skills, and abilities in content areas that cannot be adequately measured using selected-response items (Downing & Haladyna, 2006)
Performance-Based Response	Performance based tasks	To measure high-level knowledge, skills, and abilities within a complex domain.

3. Advantages and Disadvantages of AIT Items

AIT items have advantages and disadvantages. In terms of advantages, AIT items may be easier to understand, may assess higher cognitive levels, may cover a wider variety of content than traditional item types, and may be more engaging for the candidate (Sireci & Zenisky, 2006). There is some evidence within the licensure context that suggests candidates view AIT items to be more representative of real-world practice and that these items can provide additional opportunities for candidates to demonstrate their competence on more cognitively complex tasks (Lippincott, Williams, & Wilkins, 2009).

AIT items also have disadvantages. One disadvantage is that AIT items tend to be more memorable than traditional item types because of their uniqueness and images. If an item is memorable, then a candidate may share the content of the item with other candidates. This leads to a potential security problem. Additionally, AIT items may require additional response time (Dolan *et al.,* 2011) as well as development time. In particular, AIT items often require additional time to develop items, train item writers, provide item writer support, and maintain a more advanced banking/administration software. These additional requirements should be considered by any

program that is considering integrating AIT items into their exams (Institute for Credentialing Excellence, 2017).

4. Statistical Performance of AIT Items

While an abundance of literature exists to describe AIT items, much less literature is available to compare the statistical performance of these items. One study that did compare different item types occurred over 65 years ago when Frederiksen and Satter (1953) compared the difficulty of arithmetic computation items in a traditional, multiple-choice format to that of a fill-in-the-blank format. They found little difference between the difficulty of the items based on the item type. Just under 10 years later, Rimland and Zwerski (1962) also compared the performance of these two item types and found that candidates selected the distractors in well-constructed multiple-choice items at about the same frequency as candidates provided those same distractors as responses when asked the same question as a fill-in-the-blank item type. These two studies suggest that certain AIT items perform similarly to traditional, multiple-choice items.

Despite comparable statistical performances shown by these early comparisons of performance between AIT items and multiple-choice items, authors of some more recent studies have found that these very similarities can call the use of AIT items into question. On one hand, similar performance of multiple-choice and AIT items might argue against the use of the more complex AIT items. In evaluating Advanced Placement exams, Lukhele, Thissen, and Wainer (1994) concluded that constructed response items provided little information beyond that already provided by multiple-choice items. Similarly, in comparing student performance on multiple-choice and constructed response items, Hollingworth, Beard, and Proctor (2007) concluded that the extra costs and effort of scoring constructed response items might not be warranted given the similarity of the performance of the different item types. In addition, Jodoin (2003) found the AIT items he investigated (drop-and-connect and create-a-tree) provided more information than traditional, multiple-choice items, but came at the cost of seat time. This raises the question of efficiency in AIT items. Thus, it is important for any credentialing program to assess the performance of different AIT items in their own respective programs and determine whether each item type efficiently and effectively assesses the intended content.

The purpose of this study is to evaluate the performance of several AIT items on a large-scale credentialing exam. Specifically, AIT items were compared to traditional, multiple-choice items with respect to difficulty, response time, discrimination, statistical flagging, and contribution to accuracy at the cut score. It is hoped that this case study will help other credentialing programs understand some of the psychometric and pragmatic effects of integrating AIT items into an assessment program, and help to determine if the benefits of adding these items are worth the time and effort required to develop them.

5. Definitions

This study is based on a series of exams developed by an organization that are used to partially fulfill licensure requirements for a profession. Historically, this organization has used three dichotomously scored item types:

Standard Multiple Choice (SMC): Traditional four-option multiple-choice items.

Check-All-That-Apply (CATA): A question followed by a prompt to select between two and four responses out of six possible response options. All correct response options must be selected in order to answer the item correctly. This item type is also commonly referred to as multiple select or select all that apply.

Quantitative-Fill-in-the-Blank (QFIB): A question followed by an input box where candidates provide a numerical response to the question being asked. This is akin to a short answer item type in which only a numerical response is requested.

The latest version of this exam series includes two additional item types:

Hotspot (HS): A question followed by a drawing, photograph, diagram, or other image. To respond to this item type, candidates click on an area or object within the provided image.

Drag-and-Place (DnP): A question followed by a background drawing, photograph, diagram, or other image. Candidates are also presented with a series of design elements, or tokens. To respond to this item type, candidates select one or more of the tokens and place them onto the background image. This item type is also commonly referred to as drag-and-drop.

The percentage of items that were assigned to each item type varied by exam, i.e., each exam had different test specifications and the content of some exams lent itself better to certain item types than others. More details about the item type distribution across the different exams is provided in the next section.

6. Study Design

6.1 Item Types

Four AITs were compared to SMC items: CATA, QFIB, HS, and DnP. As noted previously, the SMC items in this study consisted of four options with one and only one correct answer. CATA items consisted of six options; a candidate had to select all of the correct options (between two and four) to receive credit for the item. For all exams in this study, the item informed the candidate how many options to select. QFIB items required a candidate to enter a numerical response to the item. On-screen calculators were provided to the candidates. HS items presented an image, such as a drawing or photograph, and required the candidates to click on an area of the image that correctly answered the question. DnP items had a background image and moveable elements (i.e., tokens). Candidates selected multiple tokens and dragged them into the correct places on the background image. For both HS and DnP items, item writers designated distractor regions[1]; however, a candidate could select any area within the diagram. As such, it was possible for a candidate to incorrectly answer a HS or DnP item by selecting an incorrect region that was not pre-designated as a distractor region. This is explained in more detail in the distractor analysis section of this paper.

6.2 Data

The data included in this analysis came from six exams administered nationwide and required by all 50 states' licensure programs[2]. The data included four to six pre-equated forms of each exam administered to candidates from November 1, 2016 through December 31, 2017, inclusive. Both scored and unscored items were included in the analyses. Table 2 lists the total number of items on each form and the sample size for each form.

Table 2. Total Number of Items per Form and Candidates Completing Each Form of Each Exam Analyzed

Exam	# Items/ Form	Form 1	Form 2	Form 3	Form 4	Form 5	Form 6
1	95	419	416	419	192	N/A	N/A
2	95	576	576	579	185	185	N/A
3	80	840	839	843	241	239	242
4	120	1140	1139	1141	202	202	201
5	95	544	543	545	186	185	N/A
6	120	1461	1463	1459	404	403	203

Each of the exams consists of five item types: SMC, QFIB, DnP, CATA, and HS. The number of each type of item across all forms for each exam are listed in Table 3. SMC items accounted for 54%-60% of the items administered on any one exam, while QFIB items accounted for 6%-13%, DnP items accounted for 3%-11%, CATA items accounted for 10%-29%, and HS accounted for 2%-16% of the items. The variability in the percentage of any one item type across the exams is a result of the differences reflected in the test specifications for each exam.

Table 4 provides the mean item reliability index of each item type by exam and across exams to provide a general idea of item performance. As seen in this table, QFIB items tended to raise the reliability of the exams better than the other item types; however, as discussed in this paper, it came at the cost of time.

[1]Distractor regions were invisible to the candidate. These regions were created in order to help inform item analysis. That is, they provided information about where candidates were clicking/placing tokens when they did not place them in the scored/key area(s).

[2]The authors would like to thank the National Council of Architectural Registration Boards (NCARB) for the use of their data for this research.

Table 3. Number of Administered Items Across All Forms Within Each Exam by Item Type

Item Type	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
SMC	127	129	124	183	124	211
QFIB	13	15	15	28	28	26
DnP	18	22	7	35	9	28
CATA	35	47	62	33	46	37
HS	36	24	5	45	11	47

Table 4. Mean Item Reliability Index by Exam and Item Type

Item Type	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6	Weighted Average
SMC	0.10	0.09	0.10	0.08	0.09	0.08	0.09
QFIB	0.15	0.15	0.11	0.10	0.14	0.09	0.12
DnP	0.12	0.09	0.13	0.08	0.06	0.10	0.09
CATA	0.13	0.11	0.10	0.09	0.10	0.08	0.10
HS	0.10	0.10	0.09	0.08	0.07	0.08	0.09

7. Method

Several strategies were used to determine the effectiveness of AIT items compared to SMC items. First, item difficulty (p-values), item response time, and item-total score correlations (ISC) were compared. Then, the number of statistically flagged items were compared. A kernel density plot was also constructed to compare how well the different item types targeted the minimally qualified candidate. Finally, a distractor analysis for the DnP and HS items was completed to determine how well the pre-designated distractor regions captured the candidates' responses.

8. Item p-values, Time, and ISC Comparisons

Box and whisker plots (boxplots) were constructed to compare item p-values, median item response time, and ISCs among the five item types. Boxplots were used because the number of items analyzed within each item type and exam was small; boxplots are a simple and useful method for portraying the general shape of the distribution of a variable. Specifically, boxplots show a variable's central tendency (median value) and variability. The variability is depicted by the length of the box and whiskers. For a vertical boxplot, the top and bottom of the box represent the first quartile (Q1) and third quartile (Q3) of the data, respectively. The difference between these two quartiles forms the interquartile range (IQR; i.e., the middle 50% of observations). The median is represented as a line within the box. The whiskers represent a distance from Q1 and Q3 to the highest and lowest observed values, respectively (as long as these values are within 1.5 × IQR from Q1 and Q3; if any value exceeds 1.5 × IQR, the whiskers end at the most extreme observed values that fall within the range). Any values that exceed the whiskers are considered outliers and displayed as individual dots on the plot.

Statistical comparison tests were purposely not used. The purpose of this study was not to detect statistically different differences or practical effects, but to conduct a preliminary study to identify any high-level trends among the different item types. Any identified trends could then be statistically compared in future studies with larger sample sizes.

9. Statistical Item Flagging

When items are analyzed, they can be statistically flagged based on their performance. A statistical flag does not necessarily mean there is a problem with an item; instead, it means that the item should be subjected to additional scrutiny before re-use. Therefore, a flagged item should be reviewed by subject matter experts prior to use on a future form. If a flagged item is found to have a content problem, it is either corrected and re-piloted or retired from the bank. For this study, item types were analyzed by comparing the percentage of flagged items within each

exam by item type. Like the boxplot analyses, formal statistical comparison methods were purposely not performed. Instead, the goal of this comparison was to detect high-level trends that may lead to future statistical work with larger sample sizes.

The following item flagging criteria were used for this study:

Item Level Flag Based on Difficulty or ISC

- Item was too difficult: p-value < 0.20
- Item was too easy: p-value > 0.95
- ISC for item key was not statistically and significantly different from zero
- ISC for item key had a statistically significant negative correlation

Item Level Flag Based on Top Scoring Candidates (i.e., in top quintile)

- No candidate in the top quintile answered the item correctly.
- At least one candidate in the top quintile answered the item correctly AND at least 50% of candidates in the top quintile selected the same wrong answer. ("At least 50%" must be at least 10 candidates.)

Distractor Flag

- ISC of a distractor ≥ 0.10
- Percent of candidates selecting a distractor ≥ 0.50

If item was flagged for any of the above reasons, it would also be flagged as a possible miskeyed item if one of the following occurred:

- Percent of candidates selecting a distractor *X* > 0.20 *and* ISC of distractor *X* > 0.05
- ISC of key < ISC of a distractor *X* AND all of the following occur: ISC of key < -0.10, and percent of candidates selecting distractor X > 0.20, and ISC of distractor X > 0.05

10. Kernel Density Plots

Each of these exams was analyzed using the Rasch model. As such, all of the exams had a Rasch theta passing score value. An analysis was conducted to determine which item types provided the most amount of information at the Rasch passing score. This was done by graphing a kernel density plot of the Rasch item parameters and comparing the density near the Rasch passing score for each exam. A kernel density plot is a smoothed histogram where the sum of the area under the curve equals one.

The plots in this paper show the distribution of item difficulty values broken down by item type. The plots provide similar information to (sub)test information functions for each item type but without the confounding factor of varying numbers of items by type (e.g. unadjusted (sub)test information functions would likely favor SMC items because there are more of them on the exam than other item types). The peaks of the density plots display where the Rasch values are concentrated. Since item writers for the exams target the ability region of the minimally qualified candidate, the goal (from a decision consistency point of view) is for the peak of the plot to be at or near the passing score.

11. Distractor Analysis for DnP and HS Items

A final analysis compared how well the distractors for HS and DnP items performed. For DnP items, candidates dragged objects to specific locations on a background image. The scoring mechanism then matched the items moved with the positions placed on the image. If a candidate placed an object in a location that was not in a predetermined distractor region, then the scoring mechanism recorded that position as a "Z".

An example of the use of "Z" is provided in the statistical analysis of a sample item in Table 5. In this example, 61.7% of candidates correctly responded to this item by dragging "Object 2" in the item to location "A" in the background image. However, no candidates were recorded as placing the object in location "B". From the information in the distractor analysis, 38.3% of candidates placed the object somewhere other than location "A" or "B". This large portion of candidates placing the object in an unidentified area may suggest that:

- candidates were not exact enough in their placement of the object on the background image,
- the key region is overly restrictive (e.g., not enough "wiggle room"),
- the predetermined distractor region (i.e., region "B") is not plausible, or
- item writers failed to place a distractor region in an attractive location.

In this particular example, the item-total score correlation is positive for the key and negative for the distractor, as desired. In addition, candidates who answered the item correctly spent about half of the time

on the item than those that answered it incorrectly. In other words, the item is functioning well with respect to candidate performance, but the distractor area was not placed in a position that provided useful feedback.

Table 5. Sample Item with a "Z" Response

Key	option	p-value	correlation	avg. time
>	2A	0.617	0.259	74
	2B	N/A	N/A	N/A
	2Z	0.383	-0.259	146

For HS items, candidates clicked on the background image in the location that they believed was the correct answer to the question. Not all locations were recorded as possible distractors; therefore, a "Z" in the distractor analysis indicated that the candidate selected a location on the graphic that was not pre-selected as a distractor region. This, in itself, did not mean the item was poor; it simply meant that a candidate selected a location on the graphic that did not have a predetermined distractor region. In summary, the distractor analysis for DnP and HS items helps to determine how effectively item writers create key and distractor regions.

12. Results

12.1 Item Difficulty, Time, and ISC Comparisons

An analysis was performed on the subset of CATA, DnP, HS, QFIB, and SMC items to determine if any items had outlying p-values, item response times, or ISC values and to compare the performance of the different item types across the six exams.

Figure 1 compares the p-values across the exams. Across the six exams, SMC items tended to be easier item types, but not necessarily the easiest. This is observed by noting that the box of the boxplot for SMC items tended to be contained in the upper range of the p-values for all exams, while some other item types tended to have more difficult items. However, four of the divisions showed the median difficulty of the HS items to be close to or exceeding that of the SMC items. This suggests that HS items may be an easier item type for candidates compared to other item types. In all of the divisions, CATA, DnP, and QFIB items were more difficult, on average, than SMC items. In terms of the spread of item difficulty, no clear pattern indicated one item type as being more spread out across the difficulty continuum than another item type.

Figure 2 compares the median item response times across the exams. Across the six exams, QFIB items took candidates the most amount of time to complete, on average. DnP items had the second highest median time in five of the six exams. SMC items tended to take candidates the least amount of time to complete. In terms of the spread of item response time, the only conclusive pattern emerging from Figure 2 is that QFIB items tended to have the greatest range of item response times across the different item types. This is shown by the length of the whiskers on each of the boxplots. This pattern is most apparent in Exam 1, Exam 3, Exam 5, and Exam 6.

Figure 3 compares the ISC values across the exams. Across the six exams, no consistent and conclusive pattern indicated that one item type had a higher ISC, on average, than another item type. Instead, Figure 3 suggests that all of the item types had similar ISC values, as indicated by the fact that nearly all of the boxes overlap each other within each exam. The overlap is most apparent in Exam 4 and Exam 6. In terms of the spread of ISC values, SMC items consistently had wider spreads of ISC values. DnP items tended to have narrower spreads. However, these conclusions are based on the trends seen in Figure 3 and may not be decisively related to the item type.

12.2 Statistical Flagging

Figure 4 summarizes the number of flagged items by exam and item type using the flagging criteria described earlier. While the figure does not indicate the exact criteria that caused the flag, any item flagged was a result of an item performing poorly enough that it warranted further review by subject matter experts. Although there is clearly variability across the exams, SMC items had the fewest flags while DnP items often had a high percentage of flagged items. QFIB items also had a large percentage of flags in three of the six exams.

12.3 Kernel Density Plot

The results for the kernel density plot are shown in Figure 5. The black, vertical lines in these figures represent the passing score. The x-axis is the Rasch item measure, and the y-axis is the density function value. The maximum of each density function is located at the most frequently occurring Rasch value. For the purposes of these exams (i.e., credentialing), it was highly desired to have a large

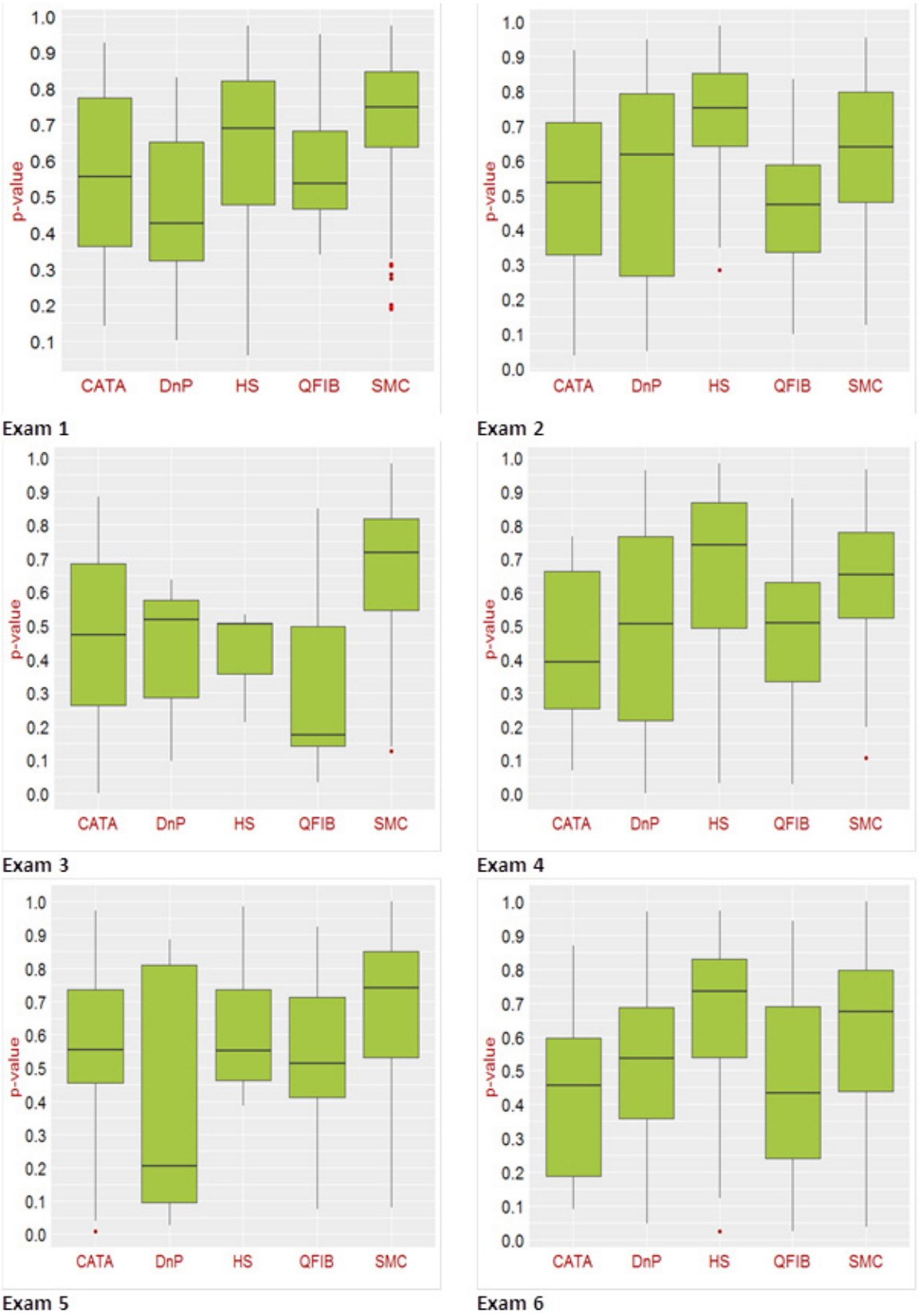

Figure 1. Comparison of p-values by exam and item type.

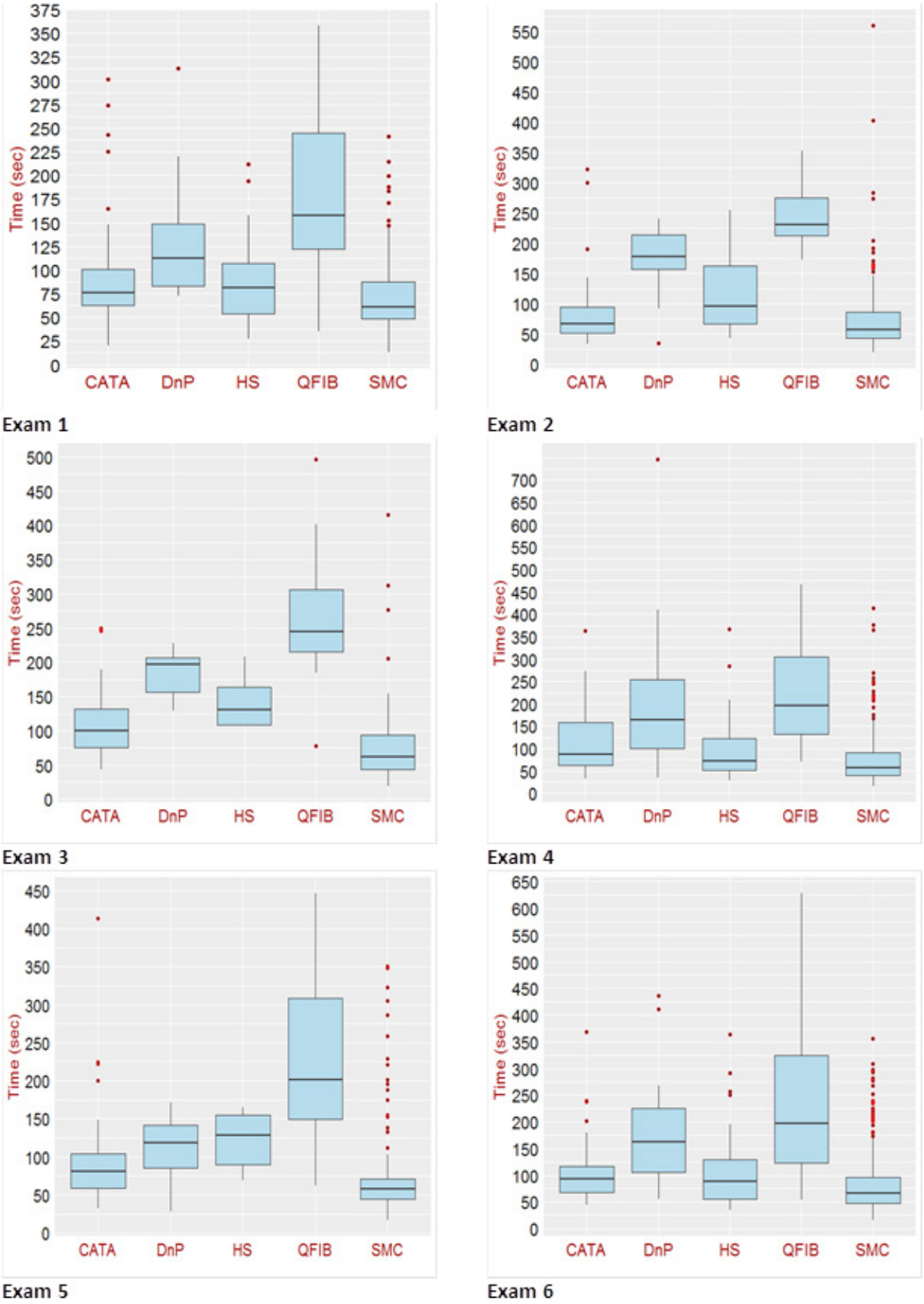

Figure 2. Comparison of item response times by exam and item type.

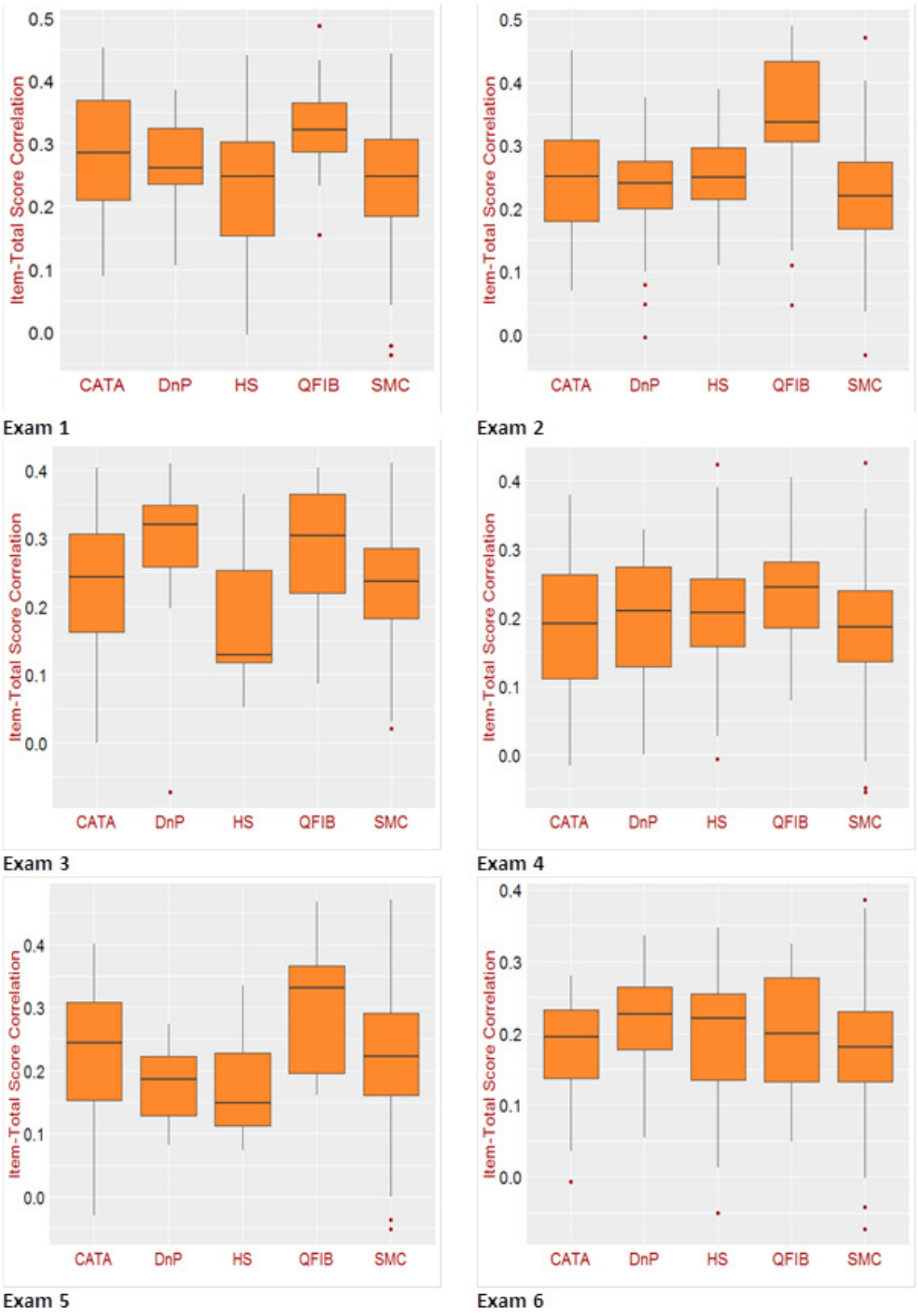

Figure 3. Comparison of ISC by exam and item type.

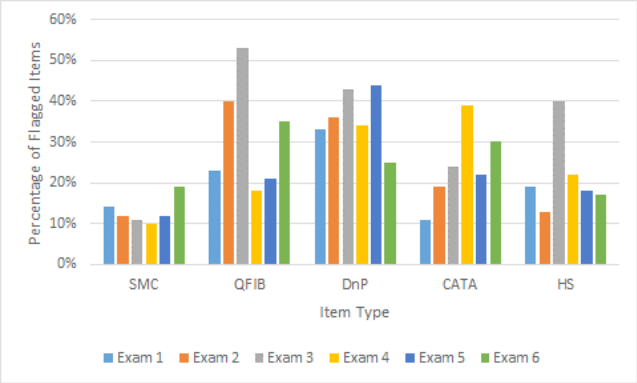

Figure 4. Percentage of flagged items by exam and item type.

number of items target the Rasch value at the passing score because this was the point at which the pass/fail decision was made.

In Figure 5, the maximum of the density function for each of the item types allows for some general trends to be noted. First, most of the QFIB items are very near the passing score in five of the six exams; the QFIB items for Exam 3 are much more difficult than the ability needed to pass the exam. The CATA items hit the ability target closer on some exams compared to others. Unlike other item types, the CATA items show bimodality in Exam 4 and Exam 6. With some exceptions, the SMC and HS items tend to be to the left of the passing score. This

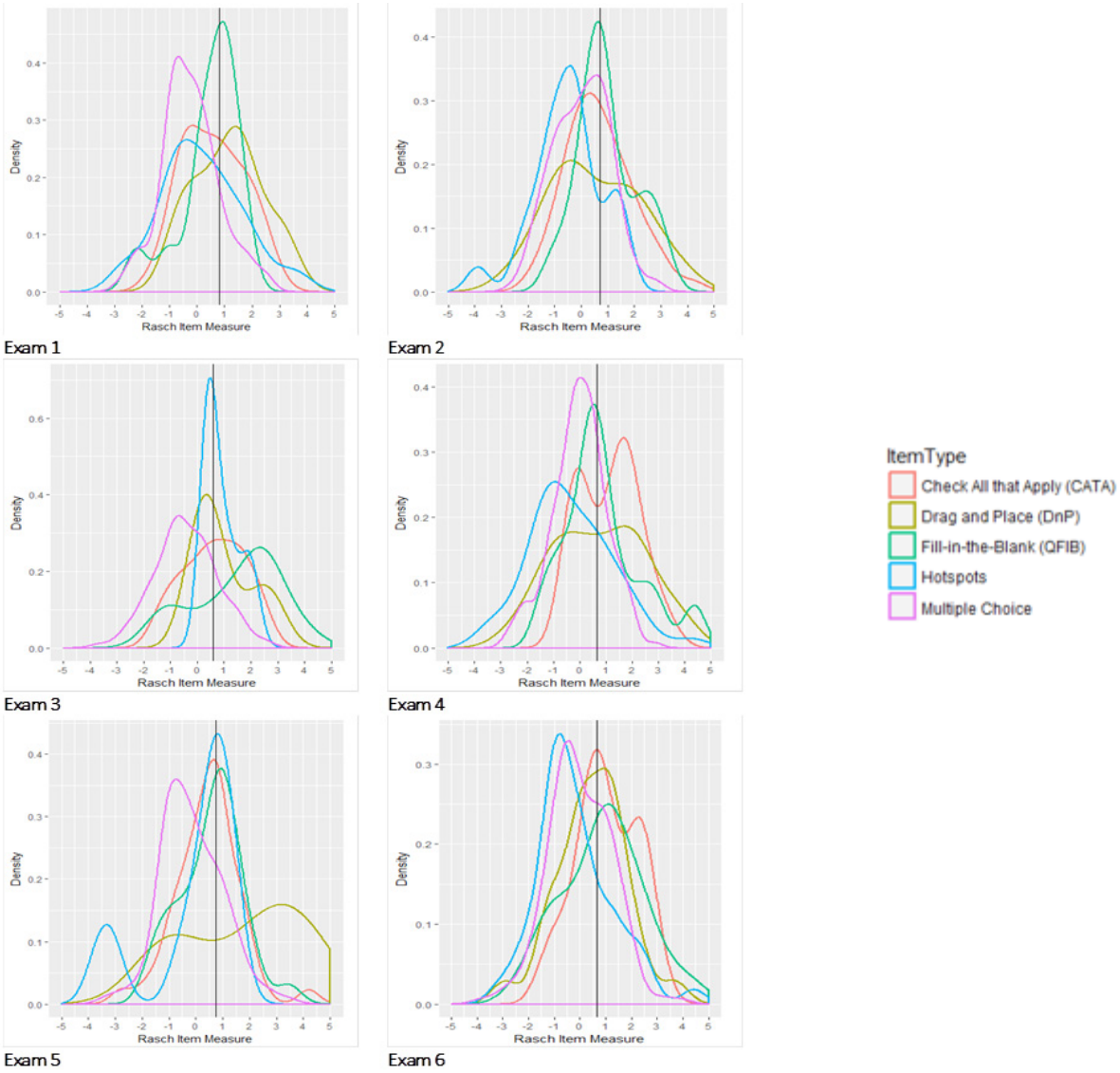

Figure 5. Kernel density plot of Rasch item parameters by exam and item type.

suggests that these items, and perhaps these item types, are easier for the minimally qualified candidate. The DnP items had the least amount of consistency across the six exams. Exam 1 and Exam 5 show the DnP items to be more difficult than the targeted ability, while Exam 2 and Exam 3 show these item types to be less difficult. Exam 4 and Exam 6 tend to center the difficulty near the targeted ability. Overall, the density plot suggests that QFIB items do well at targeting the ability level of the minimally qualified candidate for these exams while SMC and HS items tend to be less challenging. This observation is likely influenced by the item writer's ability to internalize the target ability level as it relates to different item types.

12.4 Distractor Analysis for DnP and HS Items

A distractor analysis for the AIT items that required a candidate to move objects or select a region (i.e., DnP or HS items) was analyzed in more detail. Table 6 provides a summary of the percentage of HS and DnP items in which at least 20% of candidates selected a region outside of the pre-designated distractor regions (i.e., their response was recorded as a "Z" as described earlier). While an overwhelming majority of candidates selected pre-designated distractor regions, there are still a fair number of undesignated regions selected by candidates. Although this is not a scoring problem, per se, and not necessarily an indication of a poor item, it may suggest that pre-designated regions need to be expanded, revised, or more response regions added to the item. It may indicate that item writers fail to predict where candidates would click/place tokens when answering incorrectly. When these regions are not designated as distractor regions, less information can be gathered about the "wrong" areas that candidates select.

13. Discussion

This study examined trends in AIT items by comparing item difficulty, item response time, and ISCs among SMC, CATA, QFIB, DnP, and HS items. The number of items flagged, Rasch difficulty value of the items, and a distractor analysis of DnP and HS items was also analyzed. A summary of the results is shown in the heat mapped cells of Figure 6. In general, the results indicated that SMC and HS items tended to be easier item types for candidates compared to QFIB, CATA, and DnP items. It is possible SMC items were easier because both item writers and candidates were more comfortable with this item type and that the content for these items lent themselves well to the SMC format.

The item response time analysis indicated that SMC items tended to take candidates less time to complete while QFIB items tended to take candidates more time. Again, the comfort level with SMC items may add to the candidate's ability to answer these items correctly. The longer amount of time required to answer the QFIB items is not unexpected because most of the QFIB items involved candidates completing at least one calculation.

The ISC analysis did not find any trends favoring one item type over another. Thus, from this study, it does not appear that item type has much influence over the ISC value of an item.

Across the item difficulty, item response time, and ISC comparisons, outliers were not very common. The item type that had the greatest number of outliers were SMC items in the time analysis. Since most of the SMC item response times fall within a compact range and below 100 seconds, those few items that do take longer stand out. This does not indicate that the item is poor, but if an SMC item takes considerably longer than other SMC items, it may imbalance forms if the longer items are not evenly distributed across multiple forms.

In terms of statistical flagging, DnP and QFIB items generally had more items statistically flagged compared to the other item types. SMC items had the fewest. Again, the comfort level of both item writers and candidates likely contributed to this latter finding. The high number of DnP flags may be due to technological issues related to the keyed region, but may also be due to candidates

Table 6. Percent of Items in which at Least 20% of Candidates Selected a Region Outside of the Pre-Designated Distractor Regions

Item Type	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5	Exam 6
DnP	33%	32%	0%	14%	22%	11%
HS	28%	13%	20%	20%	18%	6%

Table 4.
Overall Comparison, by Item Type and Exam

Exam	Multiple Choice (SMC)	Check All that Apply (CATA)	Fill-in-the-Blank (QFIB)	Hotspots (HS)	Drag and Place (DnP)
Median Difficulty[1]					
1	0.72	0.47	0.17	0.51	0.52
2	0.74	0.55	0.51	0.55	0.20
3	0.64	0.54	0.47	0.75	0.62
4	0.68	0.46	0.43	0.74	0.54
5	0.65	0.39	0.51	0.74	0.50
6	0.75	0.56	0.54	0.69	0.43
Median Time[2]					
1	62	100	245	131	197
2	58	81	201	128	118
3	57	67	231	97	178
4	66	93	197	88	162
5	57	87	197	73	164
6	61	76	158	81	113
Median ISC[3]					
1	0.24	0.24	0.30	0.13	0.32
2	0.22	0.24	0.33	0.15	0.19
3	0.22	0.25	0.34	0.25	0.24
4	0.18	0.19	0.20	0.22	0.23
5	0.19	0.19	0.24	0.21	0.21
6	0.25	0.29	0.32	0.25	0.26
% of Items with Statistical Flags[4]					
1	11%	24%	53%	40%	43%
2	12%	22%	21%	18%	44%
3	12%	19%	40%	13%	36%
4	19%	30%	35%	17%	25%
5	10%	39%	18%	22%	34%
6	14%	11%	23%	19%	33%

[1]Darker = more difficult (within that exam)
[2]Darker = take candidates longer (within that exam)
[3]Darker = better discriminates between high/low ability candidates (within that exam)
[4]Darker = more likely to be statistically flagged (within that exam)

Figure 6. Overall comparison, by exam and item type.

not being as comfortable with these item types. The QFIB items require candidates to manually enter a value. Historically, QFIB items tend to be difficult items because candidates do not have a forced choice decision. Instead, it is an opened-ended response. Even high ability candidates may make a simple calculation mistake or overlook a direction about how to correctly provide an answer. If enough candidates do this, the item's difficulty and ISC values will decrease.

Although QFIB items tend to be difficult, the kernel density plot analyses found that these items do a very good job of separating the minimally qualified candidates from those not qualified. Item writers appear to be overestimating the difficulty of the SMC and HS items (i.e., writing the items too easy), while they seem to still be learning how to write DnP and CATA items that target the minimally qualified ability level.

The distractor analysis for DnP and HS items indicated that these AIT items must be carefully engineered on the back end. Even if the keyed region is accurate, much information is lost if candidates select regions that are not captured in the scoring system. If all information is captured, item writers can learn the mistakes candidates are making as well as potential flaws in items. This will, in turn, help them write items that better target the minimally qualified candidate.

14. Conclusion and Considerations for Credentialing Programs

Overall, the AIT items or innovative item types of DnP, QFIB, CATA, and HS items have strong qualities. Table 7 summarizes the trends found in this study.

DnP and HS items offer a way for candidates to experience more realistic items in that they have to drag and place images into place or locate a position on an image. Based on the results in this study, selecting a location on an image is easier for candidates (and a less complex task) than dragging and placing multiple images into their proper position. The results suggested that as both candidates and item writers become more comfortable with these AIT items, these items may hit the targeted ability level more often than they currently do and have fewer flags.

QFIB items target the ability of the minimally qualified candidate well, but do take more time than other item types and tend to be statistically flagged more frequently than other item types. Both CATA and QFIB items tend to be more difficult than other item types. In general, SMC items tend to be easier to answer and tend to fall below the targeted ability level of candidates, take less time, have a wider spread of ISCs, and are flagged less often than other item types.

Deciding if it is beneficial to use one or more AITs for a credentialing exam is a programmatic decision and depends on the exam content. For example, a HS item may be well suited for a human anatomy question, but less ideal for an historical timeline question. Similarly, a QFIB item may be more appropriate for a calculation question but less appropriate for a recall question not involving a calculation. Similarly, if a multiple-choice

Table 7. Summary of Observed Trends Among the Different Item Types

Item Type	Difficulty	Item Response Time	ISC	Item Flags	Pre-designated distractor region	Targeted Ability Level
DnP	More difficult than HS and SMC	Long, but not as long as QFIB	Narrower spread than other item types	Proportionally more than other item types	Carefully selected and as comprehensive as possible	No observed trend
QFIB	More difficult than HS and SMC	Long	No observed trend	Proportionally more than CATA, HS, and SMC	N/A	Often near target
CATA	More difficult than HS and SMC	No observed trend	No observed trend	No observed trend	N/A	No observed trend
HS	Easier than most item types	No observed trend	No observed trend	No observed trend	Carefully selected and as comprehensive as possible	Often below target
SMC	Easier than most item types	Shorter than other item types	Wider spread	Proportionally fewer than other item types	N/A	Often below target

item is requiring a large amount of reading, an AIT item may potentially decrease the amount of reading and help focus the question on testing the intended objective. In general, an AIT item should be used if it can provide more opportunity to measure a candidate's ability of the intended construct than a traditional SMC item.

In addition to aligning to the exam content, security and feasibility should also be considered when using an AIT. It is important that the item type match the content, context, and purpose of the item and not just add variety to the exam. It is also important to consider the memorability of AIT items as it relates to the security and repeated use of such items. The reason for this recommendation is because DnP and HS items may be more memorable and if a traditional, multiple-choice item addresses the purpose of the question, it may be wiser to use the multiple-choice format and save the DnP and HS items for situations in which traditional, multiple-choice items do not assess the content well.

Finally, it is recommended the cognitive and psychometric properties of the AIT items be reviewed to determine if they are performing as intended or if improvements in the development of the items need to occur. From an efficiency perspective, one must weigh any improvements in validity due to the inclusion of AIT items against the extra time needed to administer them. From a security perspective, the benefit of having AIT items must be balanced with the reusability of the items. In particular, innovative item types may be more memorable for candidates. Is it worth adding the new items to an exam when the usable life of those items may be limited?

Do innovative item types really work? In summary, all of the examined item types can have psychometric and practical appeal, depending on the specific metrics one considers. While AIT items may perform well, results from this study suggest that the SMC items may still be viable alternatives. It may be best to consider using AIT items when traditional SMC items cannot assess the content in the most effective and desired way. It is also important to remember that the psychometric analysis is the evaluative portion of the item development process and not the creative process. The creativity comes from well-trained and informed content experts.

15. Limitations

The results presented here are based on a single credentialing program using data from the first months of a new version of its exams. Some of the results presented here may have been influenced by the fact that these item types were new to item writers (e.g., the relatively high proportion of statistically flagged DnP items), rather than anything inherent in the item types themselves.

These results may change over time as both item writers and candidates become more familiar with these item types. In addition, the number of DnP and HS items was relatively small. Thus, the results for these items should be interpreted cautiously.

Additionally, while we have tried to summarize broad trends in the performance of these item types, some trends exist that indicate variation occurs across exams. These six exams are all within one profession. Differences like these may be larger in exams across multiple professions.

The analyses we did were limited to classical and Rasch statistics. Further research may be done to see if these item types have different effects on, for example, the pseudo guessing parameter on a 3-PL model.

16. Further Research

The purpose of this study was to conduct a preliminary study to identify any high-level trends among the different AIT items. Future research could replicate the results of this study with exams containing a larger number of AIT items and perform statistical comparisons among the different item types to determine if the trends observed in this study continue to hold. Additional research could also be done on different AITs. Research could also investigate the efficiency of AIT items by analyzing the amount of time required to develop and administer an AIT item and compare such an item's lifetime to that of a traditional, multiple-choice item.

17. References

Dolan, R. P., Goodman, J., Strain-Seymore, E., Adams, J., & Sethuraman, S. (2011). Cognitive lab evaluation of innovative items in mathematics and English language arts assessment of elementary, middle, and high school students. [Research Report]. San Antonio, TX: Pearson.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Frederiksen, N., & Satter, G. A. (1953). The construction and validation of an arithmetical computation test. Educational and Psychological Measurement, 13(2), 209–227. https://doi.org/10.1177/001316445301300206

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. New York, NY: Routledge. https://doi.org/10.4324/9780203850381

Hollingworth, L., Beard, J. J., & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. Practical Assessment Research & Evaluation, 12(18). Available online: https://pareonline.net/getvn.asp?v=12&n=18

Institute for Credentialing Excellence. (2017). Innovative Item Types. Washington, DC: Author.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. J. Educ. Meas., 40(1), 1–15. https://doi.org/10.1111/j.1745-3984.2003.tb01093.x

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. J. Educ. Meas., 31, 234–250. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Parshall, C. G., & Harmes, J. C. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. CLEAR Exam Review, 19(2).

Rimland, B., & Zwerski, E. (1962). The use of open-end data as an aid in writing multiple-choice distracters: An evaluation with arithmetic reasoning and computation items. **J**. Appl. Psychol., *46*(1), 31–33. https://doi.org/10.1037/h0048193

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: in pursuit of improved construct representation. In S.M. Downing & T.M. Haladyna (Eds.), Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates.

Wendt, A., & Harmes, J. C. (2009). Evaluating innovative items for the NCLEX, Part I: Usability and Pilot Testing. Nurse Educator, 34(2), 56–59. https://doi.org/10.1097/NNE.0b013e3181990849