Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 3, February 2023 ISSN 1531-7714

A Method for Converting 4-Option Multiple-Choice Items to 3-Option Multiple-Choice Items Without Re-Pretesting

Amanda A. Wolkowitz, *Alpine Testing Solutions, Inc.*
Brett Foley, *Alpine Testing Solutions, Inc.*
Jared Zurn, *National Council of Architectural Registration Boards*

The purpose of this study is to introduce a method for converting scored 4-option multiple-choice (MC) items into scored 3-option MC items without re-pretesting the 3-option MC items. This study describes a six-step process for achieving this goal. Data from a professional credentialing exam was used in this study and the method was applied to 24 forms of the exam. The results found 100% accuracy in predicting the rounded passing score for all forms.

Keywords: item development, multiple choice, item format, 3-options

Introduction

There is extensive evidence in the psychometric literature showing the benefits of 3-option multiple choice (MC3) items over multiple-choice (MC) items with a greater number of response options. However, credentialing programs have been slow to adopt this format. One possible scenario that might help to explain this reluctance arises from a combination of institutional momentum and examinee test-wiseness: Consider an operational credentialing program that pre- equates its exam forms and wants to start introducing MC3 items into its exams. A reasonable course of action, and arguably best practices, would be to add MC3 items as unscored pretest items to collect data for estimating item statistics and to use in future form assembly activities. However, astute examinees may recognize the MC3 items as a new item format, guess that they are pretest items, and skip the items. Examinees who recognize that an item is a pretest item may also give less effort in responding to it compared to scored items and make the results less valid (Haladyna & Rodriguez, 2013; Pommerich & Harris, 2003). While this behavior might be unlikely, such behavior could bias the MC3 item statistics and have downstream effects on the accuracy of the pre-equating of future forms. In addition, if such a program had low-volume, then it may not be reasonable to wait up to a year to gather statistics and get pretested items approved for operational use. In an ideal situation, MC3 items could be created from existing 4-option MC (MC4) items and used as scored items on exam forms without the need for re-pretesting.

The purpose of this study is to propose and demonstrate a method in which a program may convert scored MC4 items to scored MC3 items using a combination of classical test theory (CTT) and item response theory (IRT) without re-pretesting the MC3 items first. Several recommendations for how to do this are provided. To state this upfront, it is not recommended to convert all of an exam's MC4 items into scored MC3 items without pretesting them. However, we show that it is possible to make the live conversion if only a small percentage of the items on the exam are converted and the converted items have at least one non-functioning distractor (NFD).

A program may wish to follow the proposed method so that they can honestly inform examinees that a new item type, e.g., MC3 items, will appear on

the exam as scored. For some programs, examinees may assume an item with a new format is a pretest or experimental item. Thus, it is important to communicate that this assumption is false to help ensure that examinees complete the MC3 items with as much intentionality as other items on the forms.

We begin with a summary of the literature on the performance of MC3 items, followed by an introduction to the concept of NFDs. Next, we detail our proposed methodology and apply it to a real-world testing program. Finally, we discuss the pros and cons for this method, provide recommendations for its implementation, provide practical considerations in the form of questions and answers, and consider limitations and implications for future areas of related research.

Background

Benefits of 3-Option Multiple-Choice (MC3) Items

The effect on the number of options listed for a single answer MC item is a topic that has been discussed repeatedly for over 70 years. In the 1940s, for example, Lord (1944) discussed how the number of options on a MC item changes the reliability of the test score. In the 1960s, Ebel (1967) provided formulas for estimating reliability as the number of options for an item increases. Tversky (1964) also provided a mathematical argument based on three criteria (discrimination capacity, power, and information) showing that, under certain conditions, the optimal number of options for MC items is 2.718 or 3-options.

In the late 1980s, Haladyna and Downing (1989a) published a taxonomy of MC item- writing rules. Rule 24 states, "Use as many options as are feasible; more options are desirable" (p. 40). In the same volume of Applied Measurement in Education, Haladyna and Downing offered a revision to this rule: "Develop as many functional distractors as are feasible" (Haladyna & Downing, 1989b, p. 59). They explained, "The key in distractor development is not the number of distractors but the quality of distractors" (p. 59).

In the 2010s, Haladyna & Rodriguez (2013) wrote "without any reservation" that the 3- option MC format was "superior" to the four- and five-option MC formats (p. 67). They did not state that four- or five-option MC items should not be used; instead, they reasoned that additional distractors should only be added if they are based on common errors. They continued to explain that "the creating of the fourth and fifth option seems pointless" (p. 67) because research continues to indicate the poor performance of additional distractors and item developers continue to comment on the difficulty of creating additional distractors. This sentiment was echoed in the work by Papenberg and Musch (2017) who posited that the quality of distractors is more important than the number of distractors and that additional distractors should only be added to an item if they are functional.

Much research in the 1990s, 2000s, and 2010s provide psychometric evidence that MC3 items perform just as well, if not better than, MC4 items (e.g., Baghaei & Amrahi, 2011; Bruno & Dirkzwager, 1995, Cizek, Robinson, & O'Day, 1998; Dehnad, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Mackey & Konold, 2015; Rodriguez, 2005; Rogausch, Hofer, & Krebs, 2010; Vegada, Shukla, Khilnani, Charan, & Desai, 2016; Tarrant & Ware, 2010). The psychometric benefits of MC3 items include:

1. Possible increase in exam score reliability – Exams with MC3 items have similar or higher exam score reliability than those with MC4 items. This is especially true if the number of items on the exam increases as a result of using MC3 items instead of MC4 items (Baghaei & Amrahi, 2011; Dehnad, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Grier, 1976; Haladyna & Rodriguez, 2013; Kilgour & Tayyaba, 2016; Mackey & Konold, 2015; Raymond, Stevens, & Bucak, 2018; Rogers & Harley, 1999; Schneid, Armour, Park, Yudkowsky, & Bordage, 2014; Vegada et al., 2016).
2. Little to no difference in item difficulty – MC3 items either become slightly easier or there is no significant impact on the difficulty of the item compared to MC4 items (Budesco & Nevo, 1985; Crehan, Haladyna, & Brewer, 1993; Dehnad, Nasser, & Hosseini, 2014; Loudan & Macias-Muñas, 2018; Mackey and Konold, 2015; Rodriguez, 2005; Rogaush et al., 2010; Schneid et al., 2014). It is noteworthy that if high stakes assessments start using MC3 items, then proper equating would maintain very similar pass rates and alleviate concerns about slight differences in the difficulty of the forms that may result from a

reducing the number of options on a set of items (Rogausch et al., 2010; Royal & Stockdale, 2017).

3. Possible increase in item discrimination – With the elimination of the least functioning distractor, MC3 items either have a slight increase in item discrimination or there is no significant impact on the item discrimination compared to MC4 items. In addition, the option discrimination values for MC3 items may improve when compared to the option discrimination values of the MC4 items (Baghaei & Amrahi, 2011; Cizek, Robinson, & O'Day, 1998; Dehand, Nasser, & Hosseini, 2014; Delgado & Prieto, 1998; Rodriguez, 2005; Rogausch et al., 2010; Tarrant & Ware, 2010).
4. Possible decrease in item completion time – MC3 items likely take less time to complete than MC4 items (Haladyna & Rodriguez, 2013; Schneid et al., 2014; Sidick, Barret, & Doverspike, 1994; Tarrant & Ware, 2010; Vegada et al., 2016).
5. Possible increase in content validity – MC3 items may improve the overall content validity of an exam because if there are fewer options for each item, then there could potentially be more time for more items to be on the exam; thus, more content may be covered in the same amount of time (Baghaei & Amrahi, 2011; Haladyna & Rodriguez, 2013).
6. Decrease in item development time – MC3 items take less time to develop than MC4 items (Tarrant & Ware, 2010; Vyas & Supe, 2008).
7. Possible decrease in enemy items – MC3 items provide one less option compared to MC4 items that may potentially provide a clue to another item on the test for a testwise examinee (Rogers & Harley, 1999).
8. No advantage to guessing – Described in more detail shortly, MC3 items do not increase an examinee's propensity to correctly respond to an item on a high-stakes exam. Among other reasons, this is because examinees completing high-stakes exams tend to make educated guesses and tend not to guess completely at random. Additionally, NFDs are rarely selected by examinees. Thus, an MC4 item with a throw-away option is essentially an MC3 item (Haladyna & Rodriguez, 2013; Rodriguez, 2005).

Despite the abundance of research supporting the use of MC3 items, few exam programs implement this item format. Edwards, Arthur, and Bruce (2012) collected information on the number of response options for standardized achievement tests and employment selection tests. Of the 3,051 MC items included on the exams within the 34 studies they reviewed, 53.5% were 5- option items, 43.3% were 4-option items, and only 0.9% were 3-option items.

Possible reasons for the hesitation to move to MC3 items may include that it is difficult to convince stakeholders to change the way a testing program has always delivered MC items (e.g., 4- or 5-option MC to 3-option MC), the belief that MC3 items are easier to harvest than MC4 items, and the notion that guessing makes MC3 items easier (Mackey & Konold, 2015; Baghaei & Amrahi, 2011). While the first two concerns are reasonable, they are also addressable through education and the use of a variety of item types. The belief that reducing an item from four or five options to three options makes it easier to guess the correct answer repeatedly has been shown to be a false presumption when the items are written to the examinees' targeted ability level (i.e., guessing is less likely to occur when the items are at or below the examinee's ability level) and time is not intended to be a limiting factor for the examinee (Baghaei & Amrahi, 2011; Haladyna & Rodriguez, 2013; Kilgour & Tayyaba, 2016; Rodriguez, 2005).

The quotations that follow highlight why guessing behavior should not be a limiting factor in the decision to reduce the number of options on MC items from four or five options down to three options:

> Criticism is usually made of using fewer options per item due to enhancing the probability of guessing. However, as the results of the current study revealed, multiple-choice tests, regardless of their number of options per item, would remain almost immune to the effect of guessing factor when the items are appropriately targeted for the group of test takers, and enough time has been allotted. … Although theoretically speaking the chances of getting the items right without being familiar with the construct measured in tests with three, four and five options are 33%, 25% and 20% respectively, in practice we observed that the chance factor had no influence on item difficulties across the tests (Baghaei & Amrahi, 2011, p. 207).

> The floor of a three-option item's scale is 33%, whereas with a four-option item the floor is 25% and with a five-option item the floor is 20%. Few testing programs are concerned with scores that low. Low-scoring test takers are more likely to make random guesses, and for low-scoring test takers, such variation is likely to be inconsequential … If options are implausible or non-discriminating, these four-option and five-option items are by default two- or three-option items anyway. Consequently, guessing is much overrated as a threat to validity (Haladyna & Rodriguez, 2013, p. 67).
>
> The threat of guessing and having a greater chance of a correct guess with 3-option items than with 4- or 5-option items has also not prevailed. Examinees are unlikely to engage in blind guessing, but rather educated guessing where the least plausible distractors are eliminated (Rodriguez, 2005, p. 11).

Identifying Non-Functioning Distractors (NFD)

The evidence provided thus far suggests that MC3 items are preferable to MC4 items in situations where there is an NFD. There are two common strategies for reducing the number of options in an item:

1. Randomly select a distractor to delete. Randomly selecting a distractor to delete may be sufficient for cases in which the distractors are performing equally well or, at least, all distractors are functional distractors.
2. Select the worst performing distractor to delete. The worst performing distractor may be selected by expert judgement (e.g., Dehnad, Nasser, & Hosseini, 2014; Landrum, Cashin, & Theis, 1993), empirical evidence (e.g., Kilgour & Tayyaba, 2016; Owen & Froman, 1987; Papenberg & Musch, 2017; Raymond, Stevens, & Bucak, 2018), or a combination of expert judgement and empirical evidence (e.g., Cizek et al., 1998).

For purposes of converting an MC4 item into an MC3 item when an NFD exists, randomly selecting a distractor to delete does not seem wise since the NFD may not be the option randomly selected. If the strategy chosen is to select the worst performing distractor to delete, then the task becomes identifying the criteria for selecting the NFD. If no empirical evidence exists, then the use of experts may be the only option. However, if empirical evidence has been gathered on the items from the intended population, then using this data is critical to identifying the NFD. Depending on the sample size and representativeness of the sample population used to gather the data, expert judgements used in conjunction with the data may not be needed. However, experts may be able to identify why a distractor is non-functioning. This will add to the supporting evidence and defensibility of selecting the NFD to remove.

When using empirical evidence, the method for identifying NFDs varies. Some have used the threshold of 0.05 to define an NFD, i.e., if a distractor has fewer than 5% of examinees endorsing it, then it is considered an NFD (e.g., Tarrant, Ware, & Mohammad, 2009). This threshold is problematic for two reasons. First, if an option attracts only a few low-performing examinees, then one can argue that it still plays an important role in the item (Raymond, Stevens, & Bucak, 2018; Rogausch, Hofer, & Krebs, 2010). Second, it is not uncommon for items to have p-values between 0.80 and 0.90. If an MC4 item has a p-value of 0.86, for example, then 14% of examinees who incorrectly answered the item selected one of the three distractors. By the 0.05 threshold, this means that at least two of the distractors would be considered NFDs because fewer than 5% of the examinees selected them.

Rogaush et al. (2010) recognized this issue and implemented a p-value threshold of 0.01. Raymond, Stevens, & Bucak (2018) went further and recommended that a threshold be dependent on the q-value of the item. They defined an NFD as one in which the proportion endorsing the distractor is less than 10% of those that incorrectly responded to the item, i.e., $p_{NFD} = 0.10q$ where q = 1 – p-value. Therefore, if an item has a p-value of 0.86, those distractors endorsed by fewer than 10% of those that incorrectly respond to the item (i.e., 10% × 0.14 = 1.4%) would be flagged as an NFD. If the p-value were 0.90, then distractors endorsed by fewer than 1.0% would be flagged (i.e., 10% x 0.10 = 1.0%). It is important to highlight that these flagging criteria assume that the key to the items on the exam are 100% correct and the distractors are 100% incorrect. While this seems like a reasonable assumption, in practice, this is not always an assumption one can make. Thus, additional criteria around option discrimination may be worth considering when identifying criteria for selecting the NFD.

Conversion Method

The proposed method employs six steps to identify, convert, and operationally use the MC3 version of an MC4 item:

Step 1. Determine the criteria for identifying NFDs. The criteria for identifying NFDs will be unique to each program. While there is guidance in past research to use the percent of examinees selecting a particular option, each program should review the item and option statistics for their exam and consider both these statistics and program goals when determining the appropriate NFD criteria.

Step 2. Implement the identified criteria on the current bank of items. Implement the criteria listed in Step 1 to the MC4 items with potential for conversion to an MC3 format. Any item in which at least one NFD is identified is a contender for being converted to an MC3 item. This list of potential MC3 items should be reviewed by subject matter experts to verify the appropriateness for MC3 conversions. For example, if the options for an MC4 item have parallel construction (e.g., two options discuss an increase in something and two options discuss a decrease in something), then removing a distractor results in a loss of the parallel structure of the options (which can be a useful feature).

Step 3. Update the item bank. The items approved for conversion from an MC4 to MC3 format should be updated in the exam's item bank. The update should include removing the NFD distractor from the MC4 version of the item, retiring the MC4 item or marking the MC4 item for retirement if it is currently in use, assigning a new item ID to the MC3 item, and documenting the relationship between the MC4 and MC3 versions of the item.

Step 4. Estimate the Rasch item measures for the newly converted MC3 items. The method described in this section assumes that a program pre-equates scores on an exam and that the MC items are dichotomously scored[1]. While this study uses the Rasch model for pre-equating purposes, the method described in this step could be applied with other IRT models as well as classical statistics (i.e., using p-values and total scores instead of item and person measures).

To pre-equate scores on different forms of an exam, the Rasch item measures must be known for all scored items. However, the Rasch item measures for the newly converted MC3 items are unknown at this step in the process. Consequently, they must be estimated.

When an MC4 item is converted to an MC3 item, there is no way to be certain how the change will affect the item difficulty other than the assumption that the item may become slightly easier overall. In other words, if an examinee selected an NFD in the MC4 version of the item, there is no way to know with certainty which option the examinee would select in the MC3 version of the same item. To allow for this uncertainty, consider the two most extreme situations[2]:

- Situation 1. All examinees who select the NFD in the MC4 version of the item incorrectly respond to the MC3 version of the item (i.e., score 0), and
- Situation 2. All examinees who select the NFD in the MC4 version of the item correctly respond to the MC3 version of the item (i.e., score 1).

The reality of the situation will likely fall between these two extremes. Thus, by considering both possibilities when assembling forms, there is more certainty that the change from MC4 to MC3 will not adversely affect the appropriateness of the passing scores.

To estimate the Rasch item measures for Situation 1 and Situation 2, first estimate the Rasch person measures for all examinees who completed the forms in which the MC4 version of the items appeared. During this calibration, fix all item parameters to their calibrated item bank value (if applicable) INCLUDING the MC4 items to be converted. The resulting person measures will be fixed to the calibrated measures from this point forward since a person's ability is theoretically not impacted by the item type.

[1] If a program post-equated scores on an exam, then the actual MC3 parameters could be estimated from the data and this process would not be needed.

[2] There is also no certainty that examinees who selected a non-NFD in the MC4 version of the item would continue to select the same option in the MC3 version of the same item. We have assumed that examinees who discounted the NFD when responding to the original MC4 item would not change their answer on the MC3 version. In other words, if this assumption holds, we can conclude that the only examinees who change their responses are ones who selected the NFD.

After fixing the Rasch person measures to their calibrated value, re-estimate the Rasch item measures for the converted MC3 items as follows:

- Fix the Rasch item measures to their calibrated item bank measures EXCEPT for the MC3 items.
- For Situation 1, keep the item score for each examinee who selected an NFD item as 0 for that item. For Situation 2, change the item score for each examinee who selected an NFD item to 1 for that item.
- Freely calibrate the Rasch item measures for the MC3 items for both situations[3].

Step 5. Assemble pre-equated forms and estimate a lower and upper bound for the passing score. The passing score for the forms using the newly converted MC3 items should be estimated using the Rasch measures estimated from Situation 1 and Situation 2. The Rasch item measures for the MC3 items obtained from Situation 1 provide a lower bound estimate of the passing score for each assembled form, while the Rasch item measures obtained from Situation 2 provide an upper bound estimate. Together, these two situations provide a range in which the passing score should fall[4].

Step 5a. In the special case that pass/fail decisions are needed at the time of the new form launch (i.e., data cannot be collected to verify the passing score), we recommend that MC3 items be selected in such a way that the estimated Situation 1 and Situation 2 cut scores round to the same whole-number raw cut score. This can be accomplished by selectively identifying a small number of MC3 items that have very similar Rasch parameters based on Situation 1 and Situation 2.

Step 6 (Optional, but recommended). Estimate the MC3 parameters with real data and verify the passing score. After a set of beta examinees complete each form of the exam that contain the newly converted MC3 items, the MC3 item parameters may be estimated from the real data. This calibration is completed by fixing all non-MC3 items to their calibrated item bank value and freely calibrating the MC3 items[5]. The passing score for each form can then be calculated by applying the Rasch model at the known theta passing value. This value should then be compared to the preliminary passing score established during the estimation process of steps 1 through 5.

We view Step 6 as validation, not as re-pretesting because the MC3 items will be used in the scoring of examinees. However, this step may not be possible if pass/fail decisions are needed at the time a new form is launched. In such a case, we strongly recommend programs follow steps 1 through 5a.

In the following section, we provide a walk-through of how to implement this methodology for a real-world credentialing program to show the outcome of the method in practice.

Case Study: Application of Methodology to a Real-World Credentialing Program Participants and Instruments

The data for this study came from the Architect Registration Examination® (ARE®). This is a multi-divisional, professional credentialing exam that is developed by the National Council of Architectural Registration Boards (NCARB). Passing the exam series is one of the requirements to become licensed as an architect in all 50 states as well as the District of Columbia, Guam, the Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands.

One of NCARB's goals was to convert approximately 10-20% of the existing scored MC4 items and include them as scored MC3 items on the fiscal year 2021 forms ("2021 forms"). The intent was to avoid re-pretesting all of the converted items and allow the test publisher to maintain the size of their operational item bank. Including MC3 items that were

[3] It is important to freely calibrate the Rasch item measures in both situations. While Situation 1 presents the same scoring situation as in the original data (i.e., the examinee incorrectly answered the item), the item measures should not be retained from the calibrated item bank because the item may have drifted in performance since it first appeared on a form.

[4] However, it is important to remember that this range assumes that an MC3 item is not inherently easier than an MC4 item for the examinees who did not select the NFD. To the extent that an MC3 item becomes easier for this group of examinees, this range may slightly underestimate the width of the passing score range.

[5] The Rasch person ability measures will also be freely calibrated since this is a new set of examinees completing the exam.

scored as well as newly written MC3 items that were included as embedded pretest items helped ensure examinees put forth equal effort on all items.

The items analyzed for potential conversion from MC4 to MC3 were those that appeared as either a scored or pretest item on any of the three forms of the exam administered between October 1, 2018 and September 30, 2019 ("2019 forms"). Table 1 lists the number of administrations for this time period, the number of items administered, and the percent considered for conversion. If an item was not an MC4 item (i.e., check-all-that-apply, drag and place, hotspot, or quantitative fill-in-the-blank), it was not considered for conversion.

The 2021 forms that contained the MC3 items and that were the focus of this study included items that appeared on at least one form of a divisional exam administered between December 14, 2020 and June 10, 2021. Each divisional exam administered during this period included four forms. Two of the four forms began administration on December 14, 2020. The pass/fail score decision was withheld from examinees during this period until at least 200 examinees had completed each form and analysis on the live data could be completed. The analysis included verification of the passing scores using MC3 item parameters estimated from the live data.

The second set of two forms for each divisional exam were released in February 2021. The exact date of release in February varied slightly by exam. When these latter two forms of each divisional exam were released, the exam administration favored the newly released forms. Scores were not withheld from examinees during this time period; however, verification analyses were completed on these latter two forms.

Table 2 lists the sample sizes, release date, and percent of converted and scored MC3 items by division and form. All items on the ARE are dichotomously scored. As seen in this table, approximately 7-21% of a divisional exam form contained converted MC3 items. Within a division, there was no more than a 2-item variation in the number of converted MC3 items across the four forms.

Application of the Method

The six steps described above were applied to all four forms in each of the six divisional ARE exams:

Step 1. Determine the criteria for identifying NFDs. The performance of the MC4 items administered on the 2019 forms (see Table 1) were reviewed. In particular, the option analysis was carefully reviewed to determine a reasonable threshold for identifying an NFD based on both the percent selecting a distractor and the discrimination of the distractor. The criteria were influenced by the goal of building the 2021 forms (see Table 2) with approximately 10-20% of the MC items being in an MC3 format and maintaining the same criteria for each of the six divisional exams. After careful consideration of the item and option statistics for the six exams in this program, the following criteria were set for identifying an NFD:

a) Less than 5% of the examinees who incorrectly answered the item selected the distractor, OR

b) The option-total score correlation for the distractor was positive.

Tie breaker: If an item had two NFDs, then the distractor with lower endorsement (i.e., smaller percentage of examinees selecting it) was selected as the NFD. If the NFDs had equal endorsement, then the distractor with the higher option-total score correlation was selected as the NFD.

Other programs implementing a similar conversion should determine appropriate criteria

Table 1. Items Analyzed for Potential Conversion from the 2019 Forms

Divisional Exam	# of Exam Admins.	Total N Items Administered (Scored + Pretest)	N MC4 Items Analyzed (Scored + Pretest)
A	8897	227	78
B	4520	272	123
C	6999	299	91
D	6410	287	109
E	8132	422	191
F	9991	397	137

Table 2. 2021 Forms that Included Converted MC3 Items

Divisional Exam	Form	Sample Size	Release Date*	N Total Items (Scored)	N (%) of MC3 Items (Scored)
A	1	1080	I	59	6 (10%)
	2	1089	II	59	6 (10%)
	3	1090	II	59	6 (10%)
	4	1075	I	59	6 (10%)
B	1	705	I	68	12 (18%)
	2	709	II	68	12 (18%)
	3	704	II	68	12 (18%)
	4	704	I	68	11 (16%)
C	1	802	I	68	6 (9%)
	2	804	II	68	7 (10%)
	3	798	II	68	6 (9%)
	4	805	I	68	5 (7%)
D	1	838	I	68	13 (19%)
	2	843	II	68	14 (21%)
	3	841	II	68	14 (21%)
	4	840	I	68	13 (19%)
E	1	782	I	91	10 (11%)
	2	778	I	91	11 (12%)
	3	727	II	91	11 (12%)
	4	727	II	91	10 (11%)
F	1	927	II	91	7 (8%)
	2	903	I	91	7 (8%)
	3	904	I	91	8 (9%)
	4	925	II	91	8 9%)

*I = Released December 14, 2020; II = Released February 2021

based on analysis of their own items and program goals.

Step 2. Implement the identified criteria on the current bank of items. The criteria in Step 1 were applied to the six divisional exams of the 2019 forms. Between 35% and 51% of MC4 items had at least one NFD. The exact number and percent of items are shown in Table 3. All 309 of the items were contenders for being converted to an MC3 item.

Content experts reviewed the list of potential MC4 items that could be converted to an MC3 item by removing the identified NFD. All 309 of these items were accepted for conversion to an MC3 item by removing the NFD.

Step 3. Update the item bank. All MC4 items that were converted to an MC3 item by removing the NFD were copied in the item bank and assigned a new item ID. The MC4 version of the item was either retired or marked for future retirement if it was currently in use. Notes were recorded within the item banking system to provide a record of the conversion.

Step 4. Estimate the Rasch item measures for the newly converted MC3 items. The Rasch person measures for examinees completing the 2019 forms were estimated by fixing the Rasch item measures to their known calibrated item bank measures and freely calibrating the ability measures. The Rasch person (or ability) measures were then fixed and the Rasch item measures were also fixed for all items EXCEPT the converted MC3 items.

The Rasch item measures for the converted MC3 items were re-estimated for each the two extreme situations:

- Situation 1: The item score for each examinee who selected an NFD item remained as a 0 for that item.

The Rasch item measures for the MC3 items were then freely calibrated.

- Situation 2: The item score for each examinee who selected an NFD item was changed to a 1 for that item. The Rasch item measures for the MC3 items were then freely calibrated.

For each divisional exam, a table was made that included the estimated Rasch item measure for the MC3 item under these two situations. An excerpt from one such table is shown in Table 4.

Step 5. Assemble pre-equated forms and estimate a lower and upper bound for the passing score. The passing score for the 2021 forms were estimated for Situation 1 and Situation 2. Specifically, for each divisional exam, the pre-equated forms were assembled and the lower and upper bound of the passing scores were estimated based on the values in Step 4.

Table 4 provides an example of five MC4 items that were converted to MC3 items. The estimated Rasch item measure (b) for Item 001 was -0.69 for Situation 1 and -0.74 for Situation 2. Applying the Rasch model to the known passing theta value of 0.81 (established during a previous standard setting study), Item 001 had a minimum contribution of 0.8176 points to the passing score and a maximum contribution of 0.8249 points, that is:

Situation 1: $\mathrm{P}(\text{Correct Answer}|\theta = 0.81 \; and \; b = -0.69 = \frac{e^{(.81-(-.69))}}{1+e^{(.81-(-.69))}}$

Situation 2: $\mathrm{P}(\text{Correct Answer}|\theta = 0.81 \; and \; b = -0.74 = \frac{e^{(.81-(-.74))}}{1+e^{(.81-(-.74))}}$

Thought of another way, Item 001 contributes 0.8176 points to the cut score in Situation 1 and 0.8249 points in Situation 2. The difference between these two values is approximately 0.01 raw score points. Table 5 summarizes the difference in the point contribution for each of the divisional exams. When assembling the forms, those items with the least difference between the minimum and maximum point contribution were prioritized for use. Any MC3 items not used as a scored item were set to pretest status for future use on a pretest block.

Figure 1 provides an example of the test characteristic curves (TCCs) and test information functions (TIFs) of the four forms of divisional exam

Table 3. Percent of Items with an NFD Identified by Division

Divisional Exam	N of MC4 Items Analyzed	N (%) of MC4 Items Identified as Having an NFD
A	78	27 (35%)
B	123	55 (45%)
C	91	32 (35%)
D	109	56 (51%)
E	191	85 (45%)
F	137	54 (39%)

Table 4. Examples of MC3 Rasch Item Measure Estimates (b) for 5 of the Converted Items

Item ID	Upper Bound *b* (Situation 1)	Lower Bound *b* (Situation 2)	Est. Min. Contribution to Raw Passing Score (Situation 1)	Est. Max. Contribution to Raw Passing Score (Situation 2)	Difference in Contribution to Raw Passing Score*
001	-0.69	-0.74	0.82	0.82	0.01
002	-2.06	-2.11	0.95	0.95	0.00
003	0.03	-1.28	0.69	0.89	0.20
004	-0.52	-0.52	0.79	0.79	0.00
005	0.07	0.02	0.68	0.69	0.01

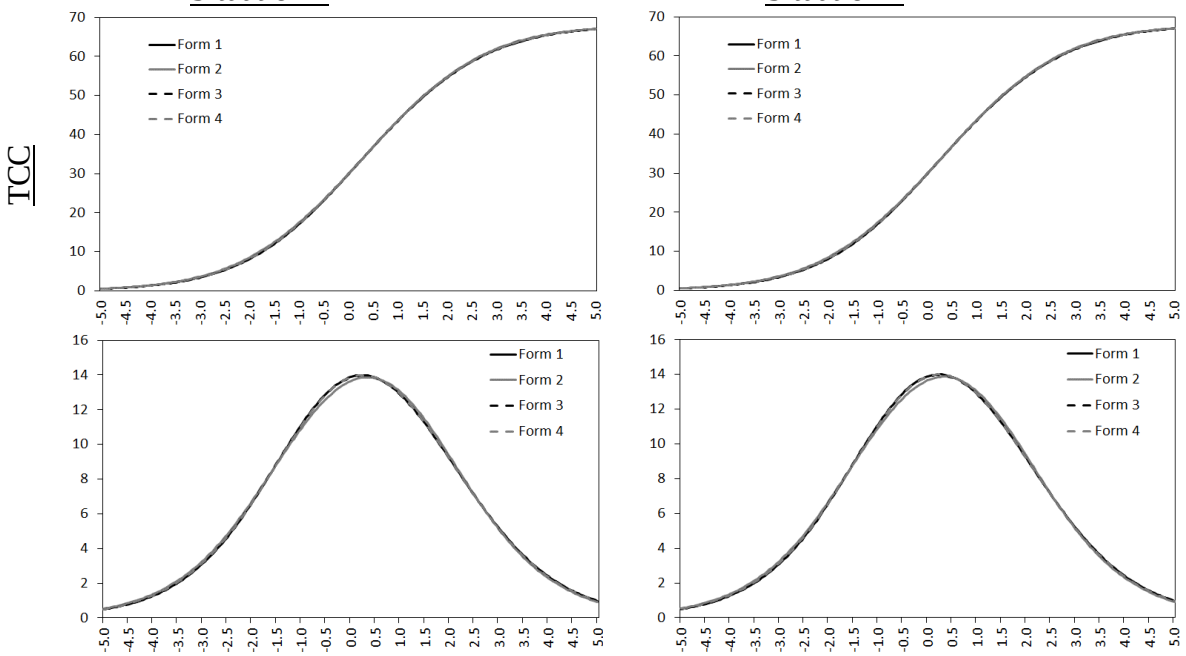
*The values in this table are rounded. Therefore, the differences in this column may not equal the difference between the displayed minimum and maximum contribution values shown in the table.

Table 5. Item Counts based on Difference in the Estimated Contribution to the Passing Score*

Difference in Contribution to Passing Score	Divisional Exam					
	A	B	C	D	E	F
< 0.010	16	38	18	29	30	24
0.010 – 0.019	4	7	5	13	8	10
0.020 – 0.029	1	1	1	3	11	4
0.030 – 0.039	1	0	0	1	4	1
0.040 – 0.049	0	0	1	0	4	0
0.050 – 0.059	0	1	0	0	1	1
0.060 – 0.069	0	1	1	1	3	1
0.070 – 0.079	1	1	0	1	1	1
0.080 – 0.089	0	2	1	0	0	0
0.090 – 0.099	0	0	0	0	0	0
≥ 0.100	4	4	5	8	21	12
TOTAL	27	55	32	56	83	54

*The difference is based on whether examinees received a 0 (Situation 1) or 1 (Situation 2) for the item in which they selected an NFD.

Figure 1. TCCs and TIFs for Exam B when MC3 Items are Estimated under Situation 1 and Situation 2



B when the Rasch measures for the MC3 items were estimated for Situation 1 (left; examinees receive 0 points on any MC3 item in which they selected an NFD in the MC4 version of the item) and Situation 2 (right; examinees receive 1 point on any MC3 item in which they selected an NFD in the MC4 version of the item). As can be seen in this figure, there was little to no impact on the coincidence of the TCCs and TIFs based on the lower and upper bound estimates of the MC3 items selected for use. The other five exams had TCCs and TIFs similar to those shown in Figure 1.

Step 6. Estimate the MC3 parameters with live data and verify the passing score. The MC3 item parameters were estimated for each form of the six divisional exams after 200 examinees had completed the form. The passing scores were then verified. For additional verification with a larger sample size and to store the MC3 item parameters into the calibrated bank for future use, the MC3 parameters were again calibrated after the June 2021 analysis. The passing scores were also rechecked at this time.

Figure 2 summarizes the results. In this figure, the triangles represent the difference between the estimated passing score based on Situation 1 (Step 5) and the final integer passing score (finalized in Step 6) for each exam form and the squares represent the difference between the estimated passing score based on Situation 2 and the final integer passing score for each exam form. The circles represent the difference between the exact passing score based on live data (Step 6) and the final rounded, integer passing score. This figure shows that all three methods round to the same integer value.

Difference Between Estimated Passing Score (Live Data, Situation 1, Situation 2) And Final Integer Passing Score

In general, the observed MC3 item parameter estimates were not within the predicted range. However, they were close. Table 6 displays the average number of logits that the MC3 item parameter estimates were from the predicted range. In this table, the average distance between the predicted range of the MC3 item parameters and the calibrated item measures based on the June 2021 data were within approximately 0.07 logits. In addition, four of the six divisional exams had a minimal, but negative average difference. This suggests that the MC3 items were easier than predicted. Shown in Table 7, the average difficulty of an MC4 item across all 24 forms decreased by approximately 0.01 (i.e., the p-value increased by 0.01) when converted to an MC3 item. This also suggests that the MC3 items were slightly easier than the MC4 version of the items. Table 7 shows that discrimination values of the MC3 items were very similar, on average, to the corresponding discrimination values of the MC4 items.

Figure 2 (shown earlier) displays the difference between the final integer passing score and the estimated passing score based on the lower bound estimate (Situation 1), live data (June 2021 analysis), and upper bound estimate (Situation 2). The forms were assembled so that the lower and upper bound estimated passing scores were approximately equal for all forms and were within approximately 0.25 points of the integer passing score. This was done to help ensure that the estimated passing score calculated from the live data would have room to deviate from the estimation and not result in the passing score being rounded to a different integer value. If the passing score did result in a different value, then NCARB was prepared to change the preliminary passing score for that form. This illustrates the importance of delaying the final scoring of exams until the passing scores can be verified.

As shown in Figure 2, all 24 forms of the 6 divisional exams had estimated passing scores within 0.50 (natural rounding rule) of the integer passing score. The greatest difference was on Form 4 of Exam B in which the estimated passing score using the June 2021 live data was approximately 0.46 points below the actual integer passing score. All remaining forms were within 0.32 points. Overall, 75% of the 24 forms had estimated passing scores within 0.10 of the final integer passing score, 90% had estimated passing scores within 0.15 points, and nearly 95% of the 24 forms had estimated passing scores within 0.20 points. A majority of the estimated passing scores in Exams B-F were underestimated by a small amount. The passing scores for Exam A, the shortest of the six exams, were consistently overestimated by a small amount.

Discussion

Assumptions

The purpose of this study was to design a method for introducing MC3 items operationally into an exam without re-pretesting them. There were several assumptions made in this study. First, it was assumed that examinees who correctly responded to an MC4 item would continue to answer the item correctly on the MC3 version of the same item. While many of these examinees would likely receive the same item score on the two versions of the item, this assumption was not tested. The second assumption was that MC3 items would not become easier than MC4 items for examinees who did not select NFDs (i.e., if they selected a functioning distractor, they were not more likely to select the answer key when an NFD was

Figure 2. Difference Between Estimated Passing Score (Live Data, Situation 1, Situation 2) And Final Integer Passing Score

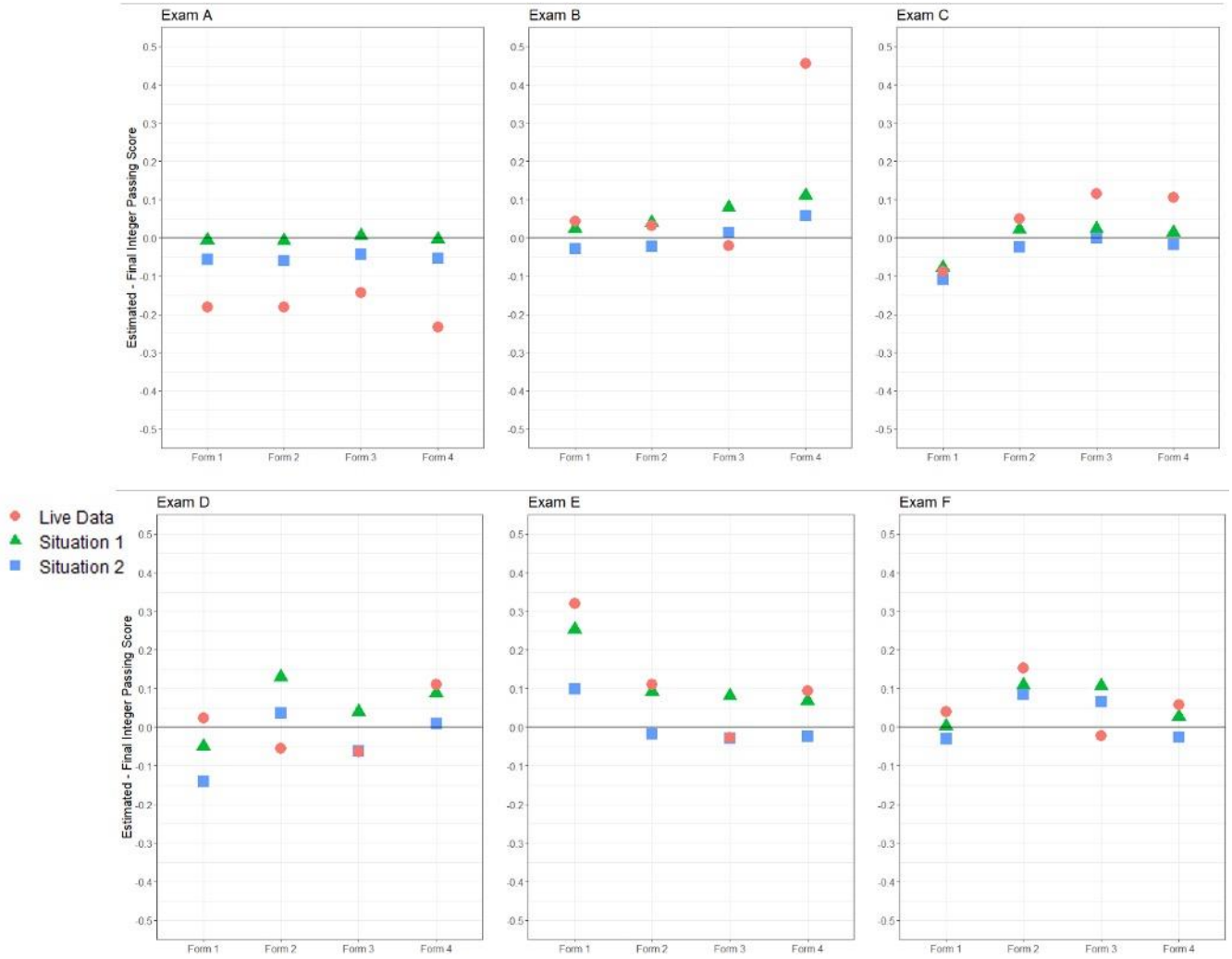


Table 6. Comparison of Predicted vs. Observed (June 2021 data) Rasch Item Measures for the MC3 Items

Divisional Exam	N items*	% of Observed Data Estimates			Average Distance between Observed and Predicted Range
		Within Predicted Range	Below Predicted Range	Above Predicted Range	
A	11	18%	36%	45%	0.05
B	27	7%	67%	26%	-0.07
C	15	7%	47%	47%	-0.01
D	33	3%	52%	45%	0.00
E	24	25%	42%	33%	-0.04
F	17	6%	53%	41%	-0.05
Overall	127	10%	51%	39%	-0.03

* The number of items in this table are less than those in Table 2 due to items overlapping across the four forms.
** The distance is the smallest difference between the June 2021 MC3 item parameter estimate and either the upper or lower bound estimate (whichever was closest).

Table 7. Comparison of Predicted vs. Observed (June 2021 data) Classical Statistics for the MC3 Items

	Average p-value		Average ISC	
Divisional Exam	**2019 MC4**	**2021 MC3**	**2019 MC4**	**2021 MC3**
A	0.66	0.67	0.19	0.18
B	0.69	0.69	0.21	0.20
C	0.70	0.70	0.21	0.19
D	0.73	0.73	0.20	0.20
E	0.64	0.66	0.17	0.17
F	0.71	0.73	0.17	0.18
Overall	0.69	0.70	0.19	0.19

* Item-total score correlation

removed). Past research has found that this is not necessarily the case, and this assumption did not appear to hold for some items in this study. That is, some items appeared to become even easier than what was anticipated in our Situation 2 (i.e., all examinees who selected the MC4 NFD responding correctly to the MC3 question). Shown in Table 6, the estimated Rasch item measure for the MC3 items using the June 2021 data tended to be less than (or easier than) the predicted ranges based on the data from the 2019 forms for five of the six divisional exams. Only Exam A, which was the shortest exam, had more item measures (i.e., 1 item) above the estimated range than below. In practical terms, however, these results suggest a minimal change in difficulty when an item is converted from an MC4 to an MC3 item. As shown in Table 7, the p-value decreased by at most 0.02. In some divisional exams, such as Exam C, there was no evidence of change in the difficulty of the MC3 version of an item compared to its MC4 version. In general, the results of this study suggest that MC3 items may be slightly easier than MC4 items, but only by a small amount. Despite it being a small change, any change should be taken into consideration when applying the methods described in this study.

A third assumption made in this study was that the other structural exam changes unrelated to the MC3 conversion occurring to the six divisional exams did not adversely affect the results. That is, at the same time NCARB was adding MC3 items to its exams, it also made some other structural changes to its exam program: proportionally shortening the number of items on the exam, increasing the relative time limit per item, introducing a new break policy, and introducing remote proctoring as a delivery option. To verify the reasonableness of this assumption, the pass rates from the 2019 forms were compared to that of the 2021 forms. The average difference in the pass rates between the 2019 and 2021 exams was approximately 4%. Individually, Exam A had an average of a 4% increase in the pass rate, Exam B had an 8% decrease, Exam C had less than a 1% decrease, Exam D has less than a 2% increase, Exam E had a 4% increase, and Exam F had an average 6% increase in the pass rate. The pass rate differences for Exams A, C, D, E, and F are reasonable for this exam program. The difference of 8% for Exam B was checked a bit further. When looking at the fiscal year 2020 data[6], the average pass rate for Exam B was only 3% higher than that of the 2021 forms. Thus, despite the changes to the exams, even the difference of 8% was considered reasonable.

A fourth assumption was that the 2019 population was similar to the 2021 population. To be sure, IRT ability/difficulty estimates tend to be robust to differences in respondent/exam ability difficulty (i.e., invariance). This assumption might not hold if the exam changes had an impact on the population or if the Covid-19 pandemic at an adverse/systematic impact on the 2021 population completing the exam. Since the content and proportionality of the exam blueprint was not changed and the pass rate differences between the fiscal year 2019 and 2021 data were reasonable, this assumption was also considered reasonable and met. If this or the previous assumption

[6] The data from the 2019 forms were used in this study instead of from the 2020 forms due to the unknown impact the Covid-19 pandemic had on the examinees completing the exam during this time.

were not met, greater changes than those observed in this study would likely have occurred.

To minimize any detrimental effects resulting from possible violations of these four assumptions, several precautionary steps were implemented. First, MC3 items selected for the forms were prioritized to include those with the least amount of difference between the lower and upper bound Rasch item measure estimates. Second, MC3 items were selected to ensure the estimated passing score under Situation 1 and Situation 2 resulted in the same integer passing score (within rounding) during the form assembly and pre-equating process. Third, the estimated upper and lower bound passing scores were as close to the same whole number as possible. These steps contributed to this study having 100% accuracy in predicting the passing score for all 24 forms across the six divisional exams.

Recommendations for Implementing the Method

The likelihood of the estimated passing score matching the passing score calculated from live data will increase if the methods as described in this paper are followed and these tips are implemented:

- *The criteria for identifying NFDs is conservative.* For example, if Program A only converts the MC4 items in which there is a distractor that no examinees select and Program B converts the MC4 items in which less than 5% of examinees selected, then Program A will likely have a more accurate estimate of the passing score than Program B. This is because Program A does not have to predict the performance of any examinees on an MC3 item while Program B does.
- *The percent of scored items converted to MC3 items is small.* Based on our results, we recommend converting fewer than 10% of the MC4 items into scored MC3 items. In this study, forms on Exams B and D had approximately 20% of the scored items converted to MC3 items and forms on Exams C and F had approximately 10% converted (see Table 2). While Figure 2 does not provide a definitive answer as to the percent of items to convert, those with only 10% converted had estimated passing scores within 0.15 of the final integer passing score across all four forms. All other exams had at least one form with an estimated passing score greater than 0.15 points from the final integer passing score. In short, the fewer items for which a program must re-estimate the Rasch parameter, the less chance of error in estimating the final integer passing score. It is also important to recognize that it is okay to have both MC4 and MC3 items on the same form of an exam.
- *Assemble forms with narrow predicted cut score ranges.* In Step 5 of the described method, the lower and upper bounds of the Rasch item measures are estimated. Using items in which the difference between the upper and lower bound Rasch measures is small helps reduce the error in the estimated passing scores. Further, the smallest differences occur for those items in which the NFD is selected by the fewest number of examinees. In this study, the maximum difference between the estimated upper and lower bound Rasch measures was 0.21. The average absolute difference was 0.04.
- *Implement a beta period.* While the methods described in this paper may be implemented without a delay in scoring, it is strongly recommended with high stakes exams to delay scoring or provide provisional scores until live data can be collected on the MC3 items and the passing score verified with the data. It is important that the final passing score be verified for fairness and validity purposes. If a program were to initially implement this method either without a delay in scoring or without provisional scoring, then the program runs the risk of the wrong passing score being applied. While this possibility can be greatly reduced by selecting very conservative criteria for selecting the NFD and minimizing the number of converted items, such an error may require retroactive action to adjust any erroneous pass/fail decisions.
- *Future MC3 items are only created through pretesting.* The methods described in this study were used to include scored MC3 items for the first time on the exam forms so that examinees would complete these items with as much intentionality as other items on the forms. Going forward, pretest blocks on these and all future forms will include MC3 items. Thus, the item parameters for these items will be collected with live data and not estimated by the described method.

While MC3 items tended to be slightly easier, on average, than the MC4 version of the same item, some items were actually more difficult. In addition, the item discrimination was not greatly impacted by the

conversion of an MC4 item into an MC3 item. While this study did not dive into the content of these items to try to understand the reason behind this occurrence, it did contribute to the success of the study. Specifically, since some item parameter ranges were over- estimates of the actual MC3 parameter (estimated form the 2021 data) and others were under- estimates, the net effect of the combination of these items on the passing score was moot to some extent.

The lingering question that may still be unanswered in a reader's mind is why did NCARB go through this entire process when they could have just delayed scoring and estimated the MC3 parameters for the items once data was available, i.e., post-equate? The answer is that NCARB wanted to assemble pre-equated forms and to eliminate, or at least minimize, any impact on examinees due to a delay in scoring. To do so required estimates of the MC3 item parameters. In addition, NCARB hoped to release the second set of 2021 forms without a delay in scoring. Consequently, once the first set of forms were analyzed and the method showed 100% accuracy in predicting the integer passing score on the 12 released forms, NCARB was comfortable releasing the second set of 12 forms with immediate scoring. NCARB was prepared to address the possible situation that a predicted passing score for a form in this second set was incorrect, but the method worked with 100% accuracy. Thus, the process described in this study allowed NCARB to successfully release forms with immediate scoring that contained scored and un-pretested MC3 items.

Questions and Answers for Applying the Proposed Method in Practice

This paper has described a method to incorporate converted MC3 from MC4 items onto an exam form without pretesting them first. For NCARB, converting a proportion of existing MC4 items into MC3 items was a practical choice supported by psychometrics and programmatic goals. NCARB weighed the risks of incorporating MC3 items into the item pool via pretesting vs. not pretesting. Given their knowledge of their stakeholders, they believed the option that would yield the most valid results in their current test development cycle was to administer operational MC3 items by converting the MC4 items into MC3 as described in this study. Other organizations may not believe this is the best option for them and, instead, inform their population that MC3 items will appear on the exam (i.e., not specify if they will be scored or unscored) and initially incorporate them as pretest items only. If this were done, then the method described in this study would not be needed. However, additional analysis of the newly introduced MC3 items should be conducted to ensure candidates did not recognize them as pretest items and disregard.

For organizations that have a more urgent need to use MC3 items operationally, the method described in this study is a potential solution. An organization may have an urgency to convert to MC3 items if they need to build new forms but have an insufficient item bank (e.g., many of the available MC4 items have a problematic distractor), if an organization wants to improve the quality of an existing form by removing NFD from existing MC4 items, or if an organization is in the middle of a test development cycle or other exam changes (as was the case for NCARB).

The questions and answers below review and highlight some of the practical considerations for converting MC4 items into MC3 items without pretesting via the method described in this study..

- *If the proposed method requires either a delay in scoring or provisional scoring, why should a program use it when they could incorporate MC3 items onto a live form and then just post-equate?* A program should use this method if they want to pre-equate forms with converted MC3 items without waiting for the MC3 items to be pretested. While the predicted item parameters for the converted MC3 items may not exactly match the observed parameters estimated after data has been collected, it does allow the forms to be built with reasonably well aligned TCCs and TIFs and reasonable estimates of the classical test theory statistics. In addition, while the proposed method highly recommends provisional scoring or delaying scoring until the predicted cut score can be verified with live data – especially for high stakes exam -- if only a limited number of scored items are converted to an MC3 item on a form and more conservative criteria is applied for removing an NFD (e.g., the NFD is selected by approximately 0% of examinees), then a delay in scoring may not be needed. Verification is still recommended.

- Why is this method better than adding MC3 items to an exam, delaying scoring, and then resetting the

item bank scale by calibrating all of the items together once sufficient data is available? It is certainly best practice and simpler to integrate a new item type (MC3 or other) as pretest items on a form and then delay scoring until sufficient data is available to calibrate the items and equate or even refresh the entire item bank scale with the inclusion of the MC3 items. However, for some programs, practical constraints may make this a less than ideal option. For some programs, a delay in scoring may not be a feasible option and the benefits of using MC3 items right away may outweigh the cons of waiting until a delay in scoring can occur, e.g., during a future standard setting study. In these situations, an organization may opt to convert only one MC4 item into a scored MC3 item and leave all other MC3 items as unscored. This would allow the program to state that the new item type will appear as both scored and unscored items on the form and also greatly reduce any possibility of error in the pre-equating of the forms.

- *What criteria for selecting items to convert from MC4 to MC3 are most effective?* It depends on the program and the item bank and how many items a program hopes to convert on the first release of the exam. The more conservative the criteria chosen, the more confidence one would have in the aligned statistical performance across the original MC4 and the new MC3 items. However, as the approach becomes more conservative, a fewer number of items will be available for the conversion. The most limiting criteria would be to convert only one MC4 item to a scored MC3 item and select an item in which nearly 0% of the examinees selected a particular distractor. If no such items exist in the item bank, then making the criteria slightly less conservative by increasing the percent selecting a distractor and considering negative item-score correlations for options can be considered. The delay in scoring is recommended for high-stakes programs that employ less conservative criteria and plan to convert multiple MC4 items into scored MC3 items. The delay in scoring for at least one set of forms will help ensure that the cut score is accurate.
- *What proportion of the exam can be converted from MC4 to scored MC3 items before the cut score can no longer be accurately estimated?* It depends on the criteria set. If a very conservative set of criteria are set, e.g., an NFD is defined as one in which no examinees select the option, then a higher proportion of items can be converted from MC4 to MC3. In the NCARB example, a relatively conservative set of criteria were used and up to 21% of the items on an exam form were newly converted MC3 items. Overall, the results from this study suggest that converting up to 10% of the scored MC4 items works well.
- *Are there other ways one could estimate the Rasch parameter for the converted MC3 items?* While this was not a focus of the current study, it is certainly possible. For example, one may create a regression equation to estimate the Rasch value from the p- value of MC4 items. Then, apply equation to estimate the MC3 item parameters from the estimated p-values of the MC3 items (after applying assumptions about which option candidates who selected an NFD would select). In addition, if an MC4 item was converted to an MC3 item by eliminating a distractor that was not selected by any examinee, then the Rasch value for the MC4 item would likely be a reasonable estimate for the MC3 item. Different estimation methods for the MC3 parameter is an area that could be studied in future research.
- *How does the conversion work for item pools with statistical characteristics of the item pool that differ from the example provided in this paper?* The conversion process would work the same. As described in this paper, care should be taken when selecting the criteria for selecting the NFD and a rationale for the decision should be documented. Depending on the item pool, there may end up being fewer or potentially more items available for conversion. It is the number of items that are converted and the criteria for selecting the NFDs that affects the success of the conversion. As previously stated, NFDs that are selected by 0% of the population and exams that only select 1 or 2 newly converted MC3 items to include as scored items on the exam will be effective regardless of the characteristics of the item bank.
- *Would this method work for converting other item types, such as check all that apply and 5-option MC items, to MC3 items?* Yes. As has been described above, care has to be taken when defining the NFD and deciding how many of such converted items will be included in the scored item section of a form. However, the general process described in this

paper should work. Given that check all that apply and 5-option MC items are quite different than MC3 items, the Rasch parameters may change more than they would for converting an MC4 item to an MC3 item. This is an area for future research.

- *When in the test development cycle is the best time to convert MC4 items to MC3 items?* Ideally, introducing a new item type would occur during the item development stage and prior to a standard setting study. However, given that some programs only conduct standard setting studies every 5-7 years (depending on the industry) and that there are multiple benefits to MC3 items, some programs may not want to wait until the next standard setting study to introduce MC3 items into their exams. Thus, implementing a method such as the one presented in this paper may provide a feasible option.
- *What other considerations should be taken if converting MC4 to MC3 items?* Stakeholders need to be informed of any changes to the test specifications, which includes the inclusion of MC3 items. In addition, there is test development costs in converting items from MC4 to MC3 items as well as potential error associated with updating the item bank. For the example provided in this paper, all MC4 items were copied and assigned a new item ID before deleting the NFD to make it an MC3 item. This was done for each converted item – one at a time. Thus, it was time consuming for those banks in which multiple items were converted. Care was also taken to document the MC3 item ID and the corresponding MC4 item ID from which it was converted. This all contributes to test development cost.

Limitations and Areas for Further Study

The limitations of this study are in the assumptions applied to the NCARB example. The greatest assumption is that the performance of an MC4 item on the 2019 form would only be impacted by the conversion of the item into an MC3 item. However, there were multiple other changes occurring to the exam at the same time. While this is certainly a limitation to the study, the differences did not seem to adversely impact the pass rate of the exam. In addition, the Board of the exam program reviewed and compared the results of the passing scores to historical data. The Board found the pass rates to be in-line with historical pass rates and approved the passing scores for each form. Examinees also were also informed of the changes, including the inclusion of MC3 items on the exam. The test publisher for these exams collects examinee feedback on each exam delivery with a post exam survey as well as a via a moderated online community. No examinees expressed concern through either channel over experiencing both MC4 and MC3 items during the same exam administration.

The other limitation to this study is that the proposed method for converting MC4 items into scored MC3 items only includes results from one exam program. Additional studies following a similar approach to that described in this study would be beneficial.

Additional research could also expand upon the research provided in this study. For example, this method could be implemented on 5-option MC items to reduce them to either 3- or 4-option items. This method may also work to reduce the number of options on a check-all-that- apply item. Future research could also determine a possible adjustment factor for the slight, but noticeable excess decrease in item difficulty resulting from removing one option in an item. There is also a research opportunity to determine if the criteria for identifying NFD, such as the ones applied in this study, are sufficient to be applied universally to other programs. Similarly, further research could explore if there is a universally recommended percentage of items on a form that a program should convert to the MC3 format.

Finally, future research could explore other methods for estimating the Rasch item measures. In this study, step 4 involved estimating the Rasch item measure for the newly converted MC3 items. This was done by fixing the Rasch item measures to their known calibrated item bank measures and freely calibrating the ability measures. These ability measures were then fixed during the estimation of the Rasch item measures. It may, for example, be more effective to estimate person ability without the items that are to be converted. Again, this is an area for further research that could improve the proposed method.

Conclusions

The purpose of this study was to introduce a method for converting scored MC4 items into scored MC3 items without re-pretesting the MC3 items first. The method used in this study included 6 steps: Determine the criteria for identifying the NFDs, implement the criteria, have experts review the items that may be converted, update the item bank with the converted items, estimate the lower and upper bound range of Rasch item measures for the newly converted MC3 items, estimate the passing score for newly assembled forms using the estimated range of Rasch item measures for the MC3 items, and verify the results with live data.

These steps were applied to a program with six divisional exams and four forms within each exam. The predicted integer passing score for each of the 24 forms matched the final integer passing score calculated from live data with 100% accuracy. In general, the unrounded predicted passing scores were slightly lower than the final, unrounded passing score estimated from the live data. This suggests that the MC3 items may have been slightly easier than the MC4 version of the same item.

Overall, the methods presented did a very good job of predicting the passing score for all 24 forms of the six exams. The reasons for the success were a result of the criteria used to identify the NFD, the small percentage of converted MC4 items, the use of MC3 items with narrow predicted ranges, and informing stakeholders of the change well in advance of the change occurring.

The method presented in this paper is a viable method for converting MC4 items into scored MC3 items. While it requires more work than post-equating and does require additional test development staff time, programs may find the benefits of using MC3 items as soon as possible outweigh the negatives of waiting until a more ideal time in the test development cycle.

References

Baghaei, P. & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological test and assessment modeling, 53*(2), 197-211.

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*(6), 959–66.

Budescu, D. V. & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22*(3), 183-196.

Cizek, G. J., Robinson, K. L., & O'Day, D. M. (1998). Nonfunctioning options: A closer look. *Educational and Psychological Measurement, 58*(4), 605–11.

Crehan, K. D., Haladyna, T. M., Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement, 53*(1), 241-247.

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A comparison between three- and four-option multiple choice questions. *Procedia – Social and Behavioral Sciences, 98*, 398-403. Retrieved from www.sciencedirect.com.

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple- choice tests. *European Journal of Psychological Assessment, 14*(3), 197-201.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement, 29*(3), 565-570. doi: 10.1177/001316446902900302

Edwards, B. D., Arthur Jr, W., & Bruce, L. L. (2012). The three-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment, 20*(1), 65-81.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*(2), 109-112.

Haladyna, T. M. & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), pp. 37-50.

Haladyna, T. M. & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item- writing rules. *Applied Measurement in Education, 2*(1), pp. 51-78.

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge: New York, NY.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education, 21*, 571–585. DOI 10.1007/s10459-015-9652-7

Landrum, R. E., Cashin, J. R., Theis, K. S. (1993). More evidence in favor of three-option multiple- choice tests. *Educational and Psychological Measurement, 53*(3), 771-778.

Lord, F. M. (1977). Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement, 14*(1), 33–38.

Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. *Journal of Educational Psychology, 35*(3), 175-180. http://dx.doi.org/10.1037/h0061025

Mackey, P. & Konold, T. R. (2015). *What is the optimal number of distractors in exam items? Case Study*. Institute for Credentialing Excellence.

Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement, 47*(2), 513-522.

Papenberg, M., & Musch, J. (2017). Of small beauties and large beasts: The quality of distractors on multiple-choice tests is more important than their quantity. *Applied Measurement in Education, 30*(4), 273–286.

Pommerich, M. & Harris, D. J. (2003). *Context effects in Pretesting: Impact on item statistics and examinee scores* [Paper presentation]. Annual Meeting of the American Educational Research Association (AERA) Conference, Chicago, IL. https://files.eric.ed.gov/fulltext/ED476923.pdf

Raymond, M., Stevens, C.., & Bucak, S. D. (2019). The optimal number of options for multiple- choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education, 24*(1), 141-150.

Rodriguez, M. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practices, 24*(2), 3-13.

Rogausch, A., Hofer, R., & Krebs, R. (2010). Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: a simulation and survey. *BMC Medical Education 10*(85), 1-9. https://doi.org/10.1186/1472-6920-10-85

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to test-wiseness and internal consistency results. *Educational and Psychological Measurement, 59*(2), 234–47.

Royal, K. D. & Stockdale, M. R. (2017). The impact of 3-Option Responses to Multiple-Choice Questions on Guessing Strategies and Cut Score Determinations. *Journal of Advances in Medical Education and Professionalism, 5*(2), 84-49. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5346173/

Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Medical Education, 48*(10), 1020-1027. https://doi.org/10.1111/medu.12525

Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology, 47*(4), 829-835.

Tarrant, A. & Ware, J. (2010). A comparison of the psychometric properties of three-and four- option multiple-choice questions in nursing assessments. *Nurse Education Today, 30*(6), 539-543.

Tarrant, M., Ware, J., & Mohammad, A. M. (2009). An assessment of functioning and non- functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9*, 40.

Tversky, A. (1964). On the optimal number of alternatives of a choice point. *Journal of mathematical psychology, 1*, 386-391.

Vegada, B., Shukla, A., Khilnani, A., Charan, J., & Desai, C. (2016). Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology, 48*, 571-5. Retrieved from http://www.ijp-online.com/text.asp?2016/48/5/571/190757

Vyas, R. & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *The National Medical Journal of India, 21*(3), 130-133.

Citation:
Wolkowitz, A. A., Foley, B., & Zurn, J. (2023). A method for converting 4-option multiple-choice items to 3-option multiple-choice items without re-pretesting. *Practical Assessment, Research, & Evaluation*, *28*(3). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/3/

Corresponding Author:
Brett Foley
Alpine Testing Solutions

Email: brett.foley [at] alpinetesting.com