RESEARCH

Differences in Multiple-Choice Questions of Opposite Stem Orientations Based on a Novel Item Quality Measure

Samuel Olusegun Adeosun, PhD
High Point University, Fred Wilson School of Pharmacy, High Point, North Carolina
Submitted October 18, 2021; accepted April 13, 2022; published February 2023.

Objective. To determine whether there are differences in the performance and quality of multiple-choice items with opposite stem orientations (positive or negative), based on a novel item quality measure and conventional psychometric parameters.

Methods. A retrospective study was conducted on multiple-choice assessment items used in years two and three of pharmacy school for pharmacotherapy and related courses administered between August 2018 and December 2019. Conventional psychometric parameters (difficulty and discrimination indices), average response time, nonfunctional distractor percentage, and a novel measure of item quality of negatively worded items were compared with those of control items, namely positively worded items (n=103 each). This novel measure uses difficulty and discrimination in tandem for the decision to reject, review, or retain items in an assessment. Statistical analyses were performed on continuous and categorical variables, on the relationship between difficulty and discrimination, and on differences in correlation coefficients between positively and negatively worded items.

Results. Stem orientation was not significantly associated with the novel measure of item quality. Also, there were no significant differences between positively and negatively worded items in any of the psychometric parameters. There were significant, negative correlations between difficulty and discrimination indices in both groups, and the correlation coefficients were significantly stronger in positively versus negatively worded items.

Conclusion. Items with opposite stem orientations show no differences in the novel item quality measure nor in conventional measures of performance and quality, except in difficulty-discrimination relationships. This suggests that negatively worded items should be used when necessary, but cautiously.

Keywords: psychometrics, item analysis, negatively worded items, multiple-choice questions, assessments

INTRODUCTION

Multiple-choice questions are one of the most commonly used assessment methods in medical and pharmacy schools because of their versatility, ease of construction, and efficiency (high reliability per hour of testing). Therefore, guidelines have been published on best practices for writing multiple-choice questions. Violating one or more of these guidelines may result in flawed questions, which may have various effects on both the students’ and item performance.

A common item-writing flaw is the negative orientation of the stem. Such items are characterized by keywords, such as not, except, or false, which ask the test taker to identify the option that is wrong rather than the option that is right, as in positively worded items. Negatively worded items are often necessary when it is important for the student to know what not to do. However, there is an additional thinking stage required to answer negatively worded items compared to positively worded items, and such items introduce the risk of double negatives when answer options also include negatively worded statements. Therefore, negatively worded items are thought to increase difficulty and negatively impact test takers’ performance. Furthermore, Chiavaroli suggested that negatively worded items behave anomalously, primarily because high-performing students get those questions wrong, but the effects of negatively worded items in several studies have largely been inconclusive.

Limitations of previous studies include the fact that negatively worded items have been analyzed jointly with other item-writing flaws and when they have been studied or analyzed separately, sample sizes were usually small (N=5 to N=37). In addition, previous
studies have compared the quality of positively and negatively worded items based on the conventional method using one or both of the major item analysis parameters (difficulty and discrimination) in isolation.\textsuperscript{3,4,7,12,14-17} Difficulty is the proportion of examination takers who get an item right, ranging from 0 to 1, representing the hardest to easiest questions, respectively.\textsuperscript{18,19} Discrimination measures how well an item differentiates between high and low scorers. Discrimination indices are calculated as either upper minus lower (U-L) or point biserial, using either a percentage of the top and bottom scorers or all test takers, respectively. Both U-L and point biserial discrimination indices are interpreted the same way and range from -1 to +1.\textsuperscript{18-21} However, because of the complex relationship between difficulty and discrimination,\textsuperscript{18,22,23} experts have recommended using both parameters in tandem rather than in isolation to gauge the quality of items in an assessment.\textsuperscript{19,24}

Given that the quality, validity, and reliability of assessments depend on the quality of items,\textsuperscript{25-28} the aim of this study was to test the null hypothesis that negatively worded items are not different from positively worded items. We used a novel measure of item quality that has not been previously used to address this question, along with the conventional psychometric parameters (difficulty and discrimination). This novel measure uses the coordinates of an item in a difficulty-discrimination matrix to inform the recommendation to either reject, review, or retain an item within an assessment. We also explored other potentially distinguishing features of negatively worded items, including average response time, distractor functionality, and correlation between difficulty and discrimination.

**METHODS**

This was a retrospective analysis of items used in summative assessments at the High Point University Fred Wilson School of Pharmacy. Data were collected in July 2021 following institutional review board approval as an exempt study. All pharmacotherapy courses and the companion integrated pharmaceutical sciences courses taken in pharmacy students’ second and third years were included. These course series were selected because they represent core components of the didactic curriculum, make up a large group of related courses from which an adequate sample size of negatively worded items were obtainable, and their assessments are primarily based on multiple-choice questions. We included only items used in midsemester and final examinations administered between August 2018 and December 2019, before the disruption due to the COVID-19 pandemic, when all lectures and examinations were still conducted live and in person.

The school of pharmacy’s ExamSoft database (ExamSoft Worldwide LLC) was searched using the negatively worded item keywords except, false, and not,\textsuperscript{3,7,8} one at a time, within the specified courses and date range. Inclusion criteria for the items included having a negatively worded multiple-choice question stem, one correct answer (key), and at least three distractors. Bonus items or items for which all test takers were given credit were excluded because they might have been considered anomalous.\textsuperscript{24} Items with multiple correct options were also excluded because of interference with distractor analysis. An item analysis report was generated for the most recent assessment in which each item was used. For each included negatively worded item, item analysis data were collected into a data collection sheet in Microsoft Excel. Average response time and distractor analysis data, including number of options, number of nonfunctional distractors (ie, distractors with percentage selection <5%), and at least three distractors. Bonus items or items with multiple correct options were also collected. Lastly, assessment data were recorded. The completed data collection sheet was cross-checked with the item analysis reports to identify and correct any data entry errors.

Control items (positively worded items) were selected for each negatively worded item during the data collection. A control item was the multiple-choice question item nearest to the negatively worded item in the item analysis report if the item was not a negatively worded item, was not a bonus/excluded item, was not already selected as a control item, and had four or more options. While this approach does not necessarily guarantee a paired match for each negatively worded item, it was used to ensure that control items were comparable. The order in which items appeared in the report was such that adjacent items addressed the same course content and were written by the same writer.

For each assessment, the percentage of negatively worded multiple-choice questions was calculated as follows: (number of negatively worded items × 100%) ÷ (number of items on the assessment − number of non–multiple-choice question items on the assessment). The percentage of nonfunctional distractors was calculated as follows: (number of nonfunctional distractors × 100%) ÷ (number of options − 1). A lower percentage of nonfunctional distractors is better.\textsuperscript{22,29}

The spread in values may mask differences when psychometric parameters are analyzed as continuous variables. Therefore, for further analysis of the conventional approach (difficulty and discrimination in isolation), difficulty was categorized as difficult/hard (<.30), good (.30-.80), or easy (> .80),\textsuperscript{31} while discrimination (27% U-L and point biserial) indices were categorized as weak (<.20), fair (.20-.29), good (.30-.39), or very good (≥.40).\textsuperscript{24,28,32}
We also categorized the items by their percentage of non-functional distractors into high- or low-quality distractor items. Since most of the items (86%) had four options and, consequently, three distractors, we categorized items with high-quality distractors as those with zero to one (0%-33.3%) nonfunctional distractors and items with low-quality distractors as those with two to three (66.7%-100%) nonfunctional distractors.

For the novel approach, items were categorized based on difficulty and discrimination in tandem. A guideline that has been shared among the school of pharmacy’s clinical sciences faculty was adapted for this study (Table 1). These criteria originated from the Lincoln Memorial University-DeBusk College of Osteopathic Medicine and provide suggestions on keeping, reviewing, or eliminating questions within an assessment based on the difficulty and discrimination indices considered in tandem. For vocabulary consistency, we maintained the review designation, while eliminate and OK designations were renamed to reject and retain, respectively. Item quality ranked best to worst was retain > review > reject. The cutoff points of this guideline generally align with other published conventions. However, for the current study, we modified the recommendation for cell B1 (Table 1) from review to OK/retain. This is because difficulty of .3-5 falls within the ideal/good range and is above the guess rate for items with four to five options (.25 or .2, respectively). Also, a discrimination index > .3 is considered good to very good by these standards.

Statistical analysis was done in SPSS version 27 (IBM Corp). Continuous variables were assessed for normality using the Shapiro-Francia test, a variation of the Shapiro-Wilk test for sample sizes greater than 5. None of the continuous variables met the condition for normal distribution (p > .05); therefore, the Mann-Whitney test was used to compare medians of control items and negatively worded items for each of the psychometric parameters. For categorical variables, the chi-square test was used to test null hypotheses about associations/differences; eg, item quality designations (retain, review, or reject) are not different, regardless of opposite stem orientations (positive or negative). The Spearman correlation was used to test correlations between difficulty and discrimination. Lastly, to determine whether correlations between the variables for the control positively worded items were different from those of the negatively worded items, we used the Fisher r-to-z transformation method. This method is also considered appropriate for nonparametric Spearman rho correlation values. In all statistical tests, α was set at .05.

RESULTS

Initial keyword search returned 334 items (of 1903 items without keywords). After reviewing and removing false positives (eg, keywords in options rather than in stems, etc), 110 negatively worded items were identified, out of which seven items were excluded (four bonus items, two items with multiple correct options, and one item with an unusually long average response time). Table 2 shows other details of the final 206 items (103 each of negatively worded items and control positively worded items) and the 36 assessments from which they were obtained. The number of negatively worded items ranged from one to 11 (mean = 2.9, SD = 2.5) items (1%-19.6%) per assessment.

Table 1. Suggested Guidelines for Reviewing and Eliminating Question Items

<table>
<thead>
<tr>
<th>Discrimination index</th>
<th>Total percentage correct (difficulty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) 0%-30%</td>
<td>(B) 30.1%-50%</td>
</tr>
<tr>
<td>(1) &gt;0.3</td>
<td>Review</td>
</tr>
<tr>
<td>(2) 0.15-0.3</td>
<td>Discard</td>
</tr>
<tr>
<td>(3) 0-0.149</td>
<td>Discard</td>
</tr>
<tr>
<td>(4) &lt;0</td>
<td>Discard</td>
</tr>
</tbody>
</table>

The original table from Lincoln Memorial University-DeBusk College of Osteopathic Medicine was reoriented to have the total percentage correct (difficulty) in columns (labeled A-D) and the discrimination index in rows (labeled 1-4). This corresponds to having the difficulty and discrimination index (upper-lower or point biserial), plotted on the x-axis and y-axis, respectively, in a scatterplot. The exact contents of the original table are presented as is, including the representation of difficulty as percentages rather than as fractions. No other modifications were made.

b “Eligible for extra credit.”

c Represents the original recommendation of cell B1, which we recategorized as OK (renamed to retain, by our designation). See notes on original document.

d “Although negative-discriminating questions with high percentages of students answering correctly are usually not discarded for practical reasons, they should probably not be used again on future exams.”
Analysis of all psychometric parameters as continuous variables showed no significant differences between control and negatively worded items (p > .05; Table 3). Chi-square was used to test the null hypotheses that there are no associations of the stem orientation (positive or negative) with difficulty (hard, good/moderate, or easy), discrimination indices (weak, fair, good, or very good), or distractor quality (high or low). There were no significant associations between stem orientation and difficulty ($\chi^2 = 5.20, p > .05$), U-L discrimination index ($\chi^2 = 0.21, p > .05$), point biserial discrimination index ($\chi^2 = 0.25, p > .05$), or item distractor quality ($\chi^2 = 0.02, p > .05$).

Using chi-square, we then tested the null hypothesis that there is no association between stem orientation (positive or negative) and the novel measure of item quality (reject, review, or retain) (Table 1). There were no significant associations between stem orientation and item quality when using difficulty in tandem with either U-L ($\chi^2 = 2.31, p > .05$; Figure 1A) or point biserial ($\chi^2 = 3.24, p > .05$; Figure 1B) discrimination indices.

Among all 206 items, there was a significant negative correlation of U-L and point biserial discrimination with difficulty ($r = -0.774, p < .001$ and $r = -0.504, p < .001$, respectively). When testing to determine whether there were differences in correlations based on the stem orientation, the Fisher r-to-z transformation analysis showed that the difficulty-discrimination correlation coefficients in control items were significantly stronger ($p < .05$) than in negatively worded items using U-L (-0.837 vs -0.708, $z = -2.319, p = .01$) and point biserial (-0.609 vs -0.397, $z = -2.031, p = .02$) discrimination indices, respectively. The distribution of the items in the matrix and the correlation trend lines are shown in Figure 1.

### DISCUSSION

Previous studies designed to determine whether negatively worded items are different in quality compared to positively worded items have used small sample sizes and conventional measures of item quality (difficulty and discrimination in isolation) and have produced mixed results. This study utilized a larger sample size and a novel measure of item quality to address these limitations.

### Table 2. Sources of Negatively Worded Items and Descriptive Statistics of Assessments Included in the Study

<table>
<thead>
<tr>
<th>Source of Negatively Worded Items</th>
<th>2018</th>
<th>2019</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWI keywords (No. (%)&lt;br&gt;Except 6 (6) 21 (20) 27 (26)</td>
<td>4 (4) 12 (12) 16 (16)</td>
<td>12 (12) 48 (47) 60 (58)</td>
<td>22 (21) 81 (79) 103 (100)</td>
</tr>
<tr>
<td>Total 22 (21) 81 (79) 103 (100)</td>
<td>5 (5) 22 (21) 27 (26)</td>
<td>17 (17) 59 (57) 76 (74)</td>
<td>22 (21) 81 (79) 103 (100)</td>
</tr>
<tr>
<td>Course series (No. (%))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharmacotherapy 5 (5) 22 (21) 27 (26)</td>
<td>17 (17) 59 (57) 76 (74)</td>
<td>22 (21) 81 (79) 103 (100)</td>
<td></td>
</tr>
<tr>
<td>Total 22 (21) 81 (79) 103 (100)</td>
<td>5 (5) 22 (21) 27 (26)</td>
<td>17 (17) 59 (57) 76 (74)</td>
<td>22 (21) 81 (79) 103 (100)</td>
</tr>
<tr>
<td>Assessment information, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of total items per assessment 69.6 (15.3) 62.0 (21.8) 64.1 (20.3)</td>
<td>3.5 (2.0) 4.9 (4.0) 4.5 (3.6)</td>
<td>62 (10) 65 (8) 64 (8)</td>
<td>0.76 (0.08) 0.72 (0.10) 0.73 (0.10)</td>
</tr>
<tr>
<td>Abbreviations: NWI = negatively worded item; MCQ = multiple-choice question; KR20 = Kuder-Richardson 20.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Continuous Variable Analysis of Psychometric Parameters of Control and Negatively Worded Items

<table>
<thead>
<tr>
<th>Psychometric Parameters</th>
<th>Control</th>
<th>NWI</th>
<th>Mann-Whitney U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>0.87</td>
<td>0.85</td>
<td>5008.0</td>
</tr>
<tr>
<td>Upper-lower discrimination index</td>
<td>0.21</td>
<td>0.20</td>
<td>5147.5</td>
</tr>
<tr>
<td>Point biserial discrimination index</td>
<td>0.25</td>
<td>0.25</td>
<td>5291.5</td>
</tr>
<tr>
<td>Average response time, s</td>
<td>52.0</td>
<td>55.0</td>
<td>5636.5</td>
</tr>
<tr>
<td>Percent nonfunctional distractor, %</td>
<td>66.7</td>
<td>66.7</td>
<td>5023.0</td>
</tr>
</tbody>
</table>

Abbreviations: NWI = negatively worded item.

* The control and NWI columns contain the median for each of the respective psychometric parameters. No significant differences were found between control and NWIs in any of the parameters ($p > .05$; n = 103 each for control and NWIs).
The current study employed a larger sample size and a novel measure of item quality that considered difficulty and discrimination in tandem. None of the individual item analysis parameters were significantly different between control and negatively worded items. More importantly, the novel measure also showed that item quality, represented by the proportions of items to be rejected, reviewed, or retained, was not significantly different between control and negatively worded items. However, the Fisher r-to-z analysis suggests that there is a significantly stronger negative correlation between difficulty and discrimination in control versus negatively worded items.

Despite the usefulness of item analysis parameters,\textsuperscript{22,31,32} they are frequently not sensitive enough....
to the detrimental effects of negatively worded items.9 Relying solely on discrimination, despite limitations, such as its nonrelevance at both extremes of difficulty, can lead to throwing out good questions.18,21,22 Also, a wide mix of item difficulties is needed for an assessment that appropriately reflects test takers’ competencies.39 It is, therefore, reasonable to leverage the potential synergy of item difficulty and discrimination when used contextually to make decisions on the quality of items on an assessment.19 As evidence of the validity of this novel measure, variations of this approach have been used to evaluate the effect of faculty training on quality of examination items written,40 item complexity on item quality,13 impact of automatic item elimination based on item quality,26 and the relationship between distractor efficiency and item quality.41 However, to the best of our knowledge, this is the first time that this measure has been used to investigate the potential difference in quality between items of opposite stem orientation.

Although we arrived at the same conclusion using both conventional and novel approaches, the novel approach is inherently different and a more pragmatic measure of item quality. For example, if discrimination were used in isolation, 40% (83 items) of the 206 items would be considered weak (using the less stringent cutoff of discrimination <.15) and, therefore, subject to deletion/rejection, while only 1% (two items) would be rejected if the novel item quality measure were used (Figure 1A). Item deletions reduce assessment quality, validity, and reliability because of reduced content coverage and mix of item difficulties.25,26,39,42 This novel item quality measure is particularly advantageous as course content in pharmacy (like most clinical disciplines) is usually large, and assessments require a broad content coverage with many need-to-know concepts (with difficulty ≈1 and discrimination ≈0). Furthermore, given that test items are often banked for subsequent use in pharmacy school in-house assessments, this novel item quality measure also has implications for evidence-based item banking processes. Including a matrix similar to Figure 1 in examination software reports may be a helpful quick guide for instructors to review and interpret item analysis data.

Average response time is a less frequently used psychometric parameter. However, we included this parameter in the analysis because it correlates with difficulty43; is associated with difficulty, complexity, and cognitive domain44; and it may be a surrogate for test takers’ effort or motivation.45 Considering the claims that negatively worded items are more difficult because of the extra time and effort needed to correctly read and interpret those questions,10,11 average response time is an appropriate multidimensional measure that should be sensitive to stem orientation. However, average response time was not significantly different between control and negatively worded items and, therefore, further supports the difficulty results.

Also, it has been suggested that one of the reasons why negatively worded items are used is because it is easier for item writers to come up with many plausible distractors for negatively worded items compared to positively worded items.9,46 Considering that the items were written by the same set of writers, our results did not support this assumption, as there was no difference between control and negatively worded items in how well their distractors functioned. Additionally, Tarrant and colleagues30 previously showed that items with a lower percentage of nonfunctional distractors are more difficult and more discriminating.30 Therefore, since the current study showed no differences in both difficulty and discrimination, the lack of difference in the percentage of nonfunctional distractors between control and negatively worded items was not surprising.

Another study26 showed a similar negative correlation between difficulty and discrimination (Figure 1). Although this coefficient was significantly different between control and negatively worded items, the practical significance of this difference may be limited, given that negatively worded items constitute a minor proportion of most assessments. While others have reported 11%-20%,5,7,14 we found an average of 4.5% negatively worded items in our study, suggesting that in a typical assessment, the relative impact of a few negatively worded items on the assessment would be negligible. That notwithstanding, this correlation difference is further evidence for perhaps limiting the use of negatively worded items, as previously suggested.3,14,15,24 Therefore, item writers in pharmacy education should use negatively worded items when necessary,21 for example, in a question to identify a drug that should not be recommended in certain comorbid conditions (eg, a nonselective beta-adrenergic antagonist in a patient with hypertension and asthma). To avoid negatively worded item overuse, one could, in this case, use the term contraindicated in the stem, which would effectively invert the orientation to a positively worded item without needing to change the multiple-choice question format (from single- to multiple-answer format) or the distractors used. However, such necessary scenarios are not limited to drug contraindications. Guidelines suggest switching the stem to the opposite (positive) orientation,2,3 but this invariably necessitates changing the multiple-choice question format15 and/or the options.14 Given that negatively worded items are one of many types of item-writing flaws, this study provides evidence that settling for negatively worded items, which will have a neutral impact on quality, is appropriate when the alternative is another item-writing flaw. For example,
rather than asking test takers to “select all that apply,” or, worse still, including implausible distractors\textsuperscript{2,16} to invert a negatively worded item, the negatively worded item format would be the better alternative.

The current results are consistent with previous studies that also found no differences between positively and negatively worded items in either their difficulty or discrimination indices, or both in isolation (Table 3).\textsuperscript{12,16,47} This includes a recent study that used 111 negatively worded items, written by the same instructor across seven courses and seven years.\textsuperscript{48} Another study also showed no difference in the psychometric properties of negatively worded items following revisions to meet multiple-choice question writing best practices.\textsuperscript{17} However, other studies have reported differences between positively and negatively worded items ranging from lower to higher difficulty or discrimination.\textsuperscript{6-8,10} Notable caveats to those studies include negatively worded items limited to one keyword,\textsuperscript{8} lack of inferential statistics,\textsuperscript{10} limited generalizability because differences were specific to certain Bloom taxonomy levels,\textsuperscript{7} and low contribution of the stem orientation relative to the effects of other interacting factors.\textsuperscript{6} But, regardless of consistency with previous studies, considering the large sample size and perhaps a more robust item quality measure used along with the conventional measures, the current study provides a stronger body of evidence for the lack of differences in the performance and quality of negatively worded items versus positively worded items. Well-designed studies are needed to verify and further demonstrate the apparent robustness advantage of the novel method over the conventional method.

Limitations of this study include that even though control items were systematically selected to be comparable to the negatively worded items, the items were still independent and not the ideal positively worded item versions of each negatively worded item. Also, even though all questions were written by the same set of writers, and the main difference between the control and negatively worded items was the stem orientation, items of both orientations might have other item-writing flaws, albeit equally. Consequently, the combined effects of these flaws may have been enough to mask the effects of stem orientation. Another inherent limitation of item analysis, when used alone or perhaps also in tandem, is its dependence on the test takers’ cohort.\textsuperscript{3,21,24,26,41,49} Lastly, instructors’ judgement is always required, as relying on psychometric parameters alone can lead to assessments with poor validity and/or reliability.\textsuperscript{21,24} Therefore, these results should be applied with caution. Future studies should use positively worded item versions of the same negatively worded items as controls in a within-subject design.

\section*{CONCLUSION}

Bearing the limitations of this study in mind, the results suggest that negatively worded items are not different from control positively worded items in the novel and conventional measures of performance and quality, except in difficulty-discrimination relationships. Therefore, in line with previous suggestions,\textsuperscript{3,14,15,48,50} negatively worded items should be used when necessary, albeit cautiously. For several reasons, negatively worded items have been called an unnecessary threat to validity,\textsuperscript{9} but based on these results, if these “when necessary” and “cautiously” provisos are considered, negatively worded items are rather benign, except, by consensus, when double negatives are not absent.

\section*{ACKNOWLEDGEMENTS}

I wish to acknowledge the Fred Wilson School of Pharmacy Academic Affairs team (Ms Amber Belvin, who wrote the ExamSoft data collection guide; Ms Gail Strickland and Peter Gal, PharmD) for their contributions and suggestions during the study planning process. I also wish to acknowledge Courtney Bradley, PharmD, Mary Jayne Kennedy, PharmD, and Peter Gal, PharmD, for reviewing various versions of this manuscript, and for their invaluable suggestions.

\section*{REFERENCES}


16. Pham H, Besanko J, Devitt P. Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple choice question. Published online 2018. doi:10.15694/mep.2018.000225.1


