Examining Validity for the Pharmacy Affective Domain Situational Judgment Test

Teresa DeLellis, PharmD,a Jaime L. Maerten-Rivera, PhD,b Yichen Zhao, PhD,b Marwa Noureldin, PharmD, MS, PhD,a Aleda M. H. Chen, PharmD, PhD,c Sharon K. Park, PharmD, MEdd

a Manchester University, College of Pharmacy, Natural & Health Sciences, Fort Wayne, Indiana
b State University of New York at Buffalo, School of Pharmacy and Pharmaceutical Sciences, Buffalo, New York
c Cedarville University, School of Pharmacy, Cedarville, Ohio
d Notre Dame of Maryland University, School of Pharmacy, Baltimore, Maryland

Submitted October 25, 2021; accepted May 4, 2022; published March 2023.

Objective. With the integration of the affective domain in the pharmacy accreditation standards, it is important for programs to have methods for formatively assessing student competency in these areas. The objective of this study was to examine the validity evidence for a recently developed situational judgment test to assess the affective domain contained in the Accreditation Council for Pharmacy Education (ACPE) Standards 2016.

Methods. After pilot testing items in multiple pharmacy schools, a revised 15-item situational judgment test instrument was developed measuring the affective domain as it relates to the pharmacy profession. The resulting Pharmacy Affective Domain Situational Judgment Test (PAD-S) was administered to students at three institutions and internal structure validity evidence was examined using item descriptives, Cronbach alpha, and results from a Rasch measurement model.

Results. There were 559 valid responses included in the study. The items were of varying difficulty, with Rasch results confirming the hierarchy of items and suggesting that items were contributing to the measure. The Cronbach alpha was 0.70, suggesting acceptable reliability. However, the reliability items from the Rasch model were lower (0.65 and 0.66), indicating that the ranking of ability was limited, which may be due to fewer items of high difficulty.

Conclusion. Overall, the PAD-S performed well as a measure of the affective domain. The PAD-S may be a useful formative instrument as part of a comprehensive assessment plan and may be less resource intensive than other measures.

Keywords: affective domain, situational judgment test, instrument validity, Rasch analysis

INTRODUCTION

The affective domain in pharmacy education comprises skills for practice and patient care as well as personal and professional development. Thus, developing and assessing this domain are essential from the health care system and patient perspectives.1 Many pharmacy programs have incorporated the affective domain within cocurricular experiences, which are experiences outside the didactic classroom and required experiential rotations.2,3

Doctor of Pharmacy (PharmD) programs have developed a variety of ways to provide, track, and assess the affective domain. One national study of pharmacy programs reported that student reflections and self-administered surveys were the most common methods of assessment.3 While these methods are valuable, there are challenges with quantifying and evaluating data from these assignments, including the workload to evaluate reflections in a systematic and timely manner. The assessment of cocurricular activities is one of the primary concerns among pharmacy faculty across the Academy.3 There are no readily available objective and comprehensive assessment instruments, which makes it difficult to assess the affective domain and to evaluate growth in these areas throughout the curriculum. Recently, there has been increased interest and use of tools such as situational judgment tests to measure one or more elements of the affective domain, including empathy, professionalism, integrity, and problem-solving.4-6

Situational judgment tests allow responders to make judgments in situations likely to be encountered in practice,
where respondents are provided a list of possible responses and asked to identify the best course of action or approach. Responses are provided in either a multiple-choice format or rank/rating based on most or least effective approaches. Generally, situational judgment tests have demonstrated evidence of validity as measurable standardized assessments and formative assessments and have increased in use recently among health professions programs and pharmacy programs specifically.4-8

Previously, a situational judgment test was developed and its items refined from cognitive interviews with pharmacy students.9 The situational judgment test took 20-30 minutes to complete and included nine questions with four options for each item that were ranked by students. It was developed to comprehensively assess the affective domains contained in Accreditation Council for Pharmacy Education (ACPE) Standards 3 and 4, namely problem-solving, education, patient advocacy, interprofessional collaboration, cultural sensitivity, communication, self-awareness, leadership, innovation and entrepreneurship, and professionalism. However, the study did not examine internal structure validity of the situational judgment test to measure the affective domain.9 An instrument that measures competency in the affective domain while using minimal program resources would be of value for formative curriculum assessment, especially for students’ growth and progression. Therefore, the purpose of this study was to report the psychometric properties of the situational judgment test in assessing the affective domain.

METHODS

Examining the validity of a measure requires integrating various aspects of evidence collected.10 Possible sources of validity evidence include test content (relevance and representativeness of the measure, generally assessed by examining the items), response processes (examination of the process employed by an examinee to yield a response), and internal structure (psychometric analyses used to evaluate items and reliability). Figure 1 outlines the process used to develop and evaluate the validity of the situational judgment test.

In this study, a previously developed and tested situational judgment test was revised to include 15 items to measure the affective domain; the resulting test is the Pharmacy Affective Domain Situational Judgment Test (PAD-S), which is available without charge to any institution upon request. The PAD-S items mapped to each element of Standards 3 and 4 are presented in Table 1. The PAD-S was administered to first-year (P1) through fourth-year (P4) pharmacy students from three institutions (Manchester University, Cedarville University, and University at Buffalo) using the online testing platform ExamSoft (ExamSoft Worldwide LLC). The institutions represented different elements of the Academy. Manchester is a private Midwest institution with 40% of students identifying as non-White. Cedarville is a private Midwest institution; approximately 33% of students identify as non-White. University at Buffalo is a public institution located in the Northeast; approximately 57% of students identify as non-White. All students were asked to participate in the PAD-S but were given an option to withhold their results from the study data analysis. A total of 630 students completed the PAD-S, and 559 (88.7%) of those gave consent to the research and were included in the analyses. Internal structure validity evidence was examined for the 15-item PAD-S. As there is no consensus regarding situational judgment test scoring methods, two scoring strategies were explored: in the first method, one point was awarded for each correctly ordered response (zero to three points total); in the second method, one point each was given for the correctly chosen most effective and least effective selections (zero to two points total).11

As a first step in the analyses, the frequency of responses for each item was examined along with the total scores on the PAD-S and Cronbach alpha (SPSS version 25.0 [IBM Corp]). Cronbach alpha is often used to estimate internal consistency reliability, with coefficient estimates of 0.70 or above considered generally acceptable.12 Additionally, the Rasch model was used to augment Cronbach alpha by taking into consideration each student’s ability (ie, overall affective domain ability estimate) and each item’s difficulty, and by examining how students interacted with each item based on their ability.12-14 Rasch models provide reliability and internal structure validity evidence, including fit statistics and person-item maps, and estimate how well the items function together as a unidimensional measure (ie, item fit).14-16 The item fit statistics estimated are the infit and outfit mean-square (MNSQ), which measure whether the item is contributing (or helpful) to the measure; values between 0.6-1.4 are considered acceptable.16 Item difficulty estimates are provided in logits (with a mean=0 and SD=1) and estimate how challenging an item is; negative difficulty estimates are given to easier items, while positive estimates are of higher difficulty.16

Rasch analyses also provide estimates of item and person reliability as well as item and person separation. The person reliability estimate is similar to traditional test reliability estimates and can be interpreted similarly to Cronbach alpha, with values of 0.70 or higher considered acceptable. Separation gives additional information regarding reliability and can be conceptualized as the number of units on the instrument’s measurement “ruler,” with estimates less than 2.0 being of concern.16,17 Rasch analyses
Figure 1. Outline of the process used to develop and evaluate the validity of a Pharmacy Affective Domain Situational Judgment Test (PAD-S).

also provide an item-person map, which is useful as it demonstrates the match of each item’s difficulty to the estimated abilities for the respondents using logits for both estimates. This map is key to determining whether the items’ difficulty levels are aligned with the respondents’ overall abilities. Generally, the items should have spread

Table 1. Pharmacy Affective Domain Situational Judgment Test Items Mapped to the Affective Domains of ACPE Standards 3 and 4

<table>
<thead>
<tr>
<th>Standard</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>Q14</th>
<th>Q15</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Problem solving</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3.2 Education</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3 Patient advocacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4 Interprofessional collaboration</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5 Cultural sensitivity</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.6 Communication</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1 Self-awareness</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 Leadership</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3 Innovation and entrepreneurship</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4 Professionalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: PAD-S = Pharmacy Affective Domain Situational Judgment Test; ACPE = Accreditation Council for Pharmacy Education.
across all levels of the ability estimates, with more items having difficulty levels aligned with the middle range of the ability estimates. Rasch models were examined using Winsteps version 3.72.0 (Winsteps.com). This study was approved by the institutional review board of Manchester University, IN.

RESULTS

There were 559 valid responses in the study, which included 191 P1 students, 154 P2 students, 172 P3 students, and 42 P4 students from the three institutions. For scoring option 1, where one point was given for each correctly ordered response (max score = 45), the mean (SD) score was 30.10 (5.65). For scoring option 2, where one point each was given for the most effective and least effective selections (max score = 30), the mean (SD) score was 20.40 (4.26). The Cronbach alpha for both scoring options was 0.70, indicating an acceptable level of internal consistency for either method. The frequency of points received for each question are presented in Table 2. For all questions and both scoring options, students generally received at least one point, although there was variation in the points received (ie, difficulty). This provides evidence that not all items were of the same difficulty, and, therefore, Rasch analyses were appropriate to further examine student ability and item difficulty.

The infit MNSQ, outfit MNSQ, and item difficulty estimates from the Rasch model are reported in Table 2 for both scoring options. For most of the items, the infit and outfit MNSQ were within the acceptable range of 0.6-1.4 (scoring option 1 Q1 and Q2 were slightly outside of the range), suggesting that most items are contributing to the measure. The range of item difficulty for scoring option 1 was -0.97 to 0.91, and for scoring option 2 the range was -1.15 to 1.23. This suggests that both scoring options contained a range of easy (item difficulty estimates < 0), moderate (close to 0), and difficult (> 0) items, though scoring option 2 had a wider range of difficulty.

The reliability and separation estimates from the Rasch model are presented in Table 3. For both scoring options, the item reliability was 0.99 and the item separation estimates were > 10, indicating that the item difficulty hierarchy of the situational judgment test measure was confirmed. The person reliability estimates for scoring options 1 and 2 were 0.65 and 0.66, respectively (slightly below the 0.70 threshold for indicating an acceptable level of internal consistency). The person separation estimates were 1.37 and 1.38 (less than the 2.0 threshold), indicating the measure could be improved to distinguish between individuals (ie, ranking the ability of one student relative to another consistently was limited).

With both scoring options, there were very few ability estimates below -1, indicating few students with low

Table 2. Fit Statistics and Item Difficulty for Pharmacy Affective Domain Situational Judgment Test Items by Scoring Method

<table>
<thead>
<tr>
<th>PAD-S</th>
<th>Item</th>
<th>0 points</th>
<th>1 point</th>
<th>2 points</th>
<th>3 points</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>Item difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>15.9</td>
<td>31.1</td>
<td>35.2</td>
<td>17.7</td>
<td>0.15</td>
<td>1.51</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>23.6</td>
<td>17.7</td>
<td>42.4</td>
<td>16.3</td>
<td>0.14</td>
<td>1.44</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>4.3</td>
<td>9.1</td>
<td>31.1</td>
<td>55.5</td>
<td>1.07</td>
<td>-0.63</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>10.6</td>
<td>46.3</td>
<td>36.0</td>
<td>7.2</td>
<td>0.83</td>
<td>0.91</td>
<td>-0.97</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>4.7</td>
<td>5.4</td>
<td>23.8</td>
<td>66.2</td>
<td>1.14</td>
<td>-0.97</td>
<td>0.91</td>
<td>-0.86</td>
</tr>
<tr>
<td>Q6</td>
<td>3.2</td>
<td>4.7</td>
<td>32.0</td>
<td>60.1</td>
<td>0.95</td>
<td>-0.90</td>
<td>0.91</td>
<td>-0.86</td>
</tr>
<tr>
<td>Q7</td>
<td>5.4</td>
<td>5.7</td>
<td>31.8</td>
<td>57.1</td>
<td>1.05</td>
<td>-0.69</td>
<td>0.91</td>
<td>-0.75</td>
</tr>
<tr>
<td>Q8</td>
<td>15.4</td>
<td>9.8</td>
<td>33.8</td>
<td>41.0</td>
<td>1.29</td>
<td>0.06</td>
<td>0.91</td>
<td>-0.72</td>
</tr>
<tr>
<td>Q9</td>
<td>15.7</td>
<td>23.6</td>
<td>36.9</td>
<td>23.8</td>
<td>1.06</td>
<td>1.05</td>
<td>0.53</td>
<td>-0.86</td>
</tr>
<tr>
<td>Q10</td>
<td>8.9</td>
<td>21.1</td>
<td>51.2</td>
<td>18.8</td>
<td>0.83</td>
<td>0.84</td>
<td>0.37</td>
<td>-0.86</td>
</tr>
<tr>
<td>Q11</td>
<td>2.9</td>
<td>11.3</td>
<td>42.4</td>
<td>43.5</td>
<td>0.95</td>
<td>0.93</td>
<td>0.00</td>
<td>-0.86</td>
</tr>
<tr>
<td>Q12</td>
<td>4.1</td>
<td>10.0</td>
<td>63.5</td>
<td>22.4</td>
<td>0.65</td>
<td>0.63</td>
<td>0.00</td>
<td>-0.86</td>
</tr>
<tr>
<td>Q13</td>
<td>5.0</td>
<td>14.3</td>
<td>38.8</td>
<td>41.9</td>
<td>0.94</td>
<td>0.92</td>
<td>-0.23</td>
<td>-0.40</td>
</tr>
<tr>
<td>Q14</td>
<td>8.8</td>
<td>24.3</td>
<td>40.3</td>
<td>26.7</td>
<td>1.01</td>
<td>1.02</td>
<td>0.30</td>
<td>-0.22</td>
</tr>
<tr>
<td>Q15</td>
<td>6.8</td>
<td>14.3</td>
<td>56.4</td>
<td>22.5</td>
<td>0.80</td>
<td>0.78</td>
<td>0.15</td>
<td>-0.22</td>
</tr>
</tbody>
</table>

Abbreviations: PAD-S = Pharmacy Affective Domain Situational Judgment Test; MNSQ = mean-square.

* The item fit statistics estimated are the infit and outfit mean-square (MNSQ), which measure whether the item is contributing to the measure; values between 0.6-1.4 are considered acceptable. Item difficulty estimates are provided in logits (with a mean = 0 and SD = 1) and estimate how challenging an item is; negative difficulty estimates are given to easier items, while positive estimates are of higher difficulty.
ability estimates. Item difficulties were distributed between -1 logit and 1 logit, while person abilities were distributed between -1 logit and 2 logits. The items corresponded to the ability estimates for many of the students; however, the top student ability estimates were above the most difficult item, indicating that the items are not difficult enough to assess high person abilities. This inability to reliably measure and rank high performers is reflected in the low person reliability mean. The person-item maps delineating the relationship between student ability and item difficulty were reviewed and are available upon request.

**DISCUSSION**

Pharmacists with strong affective domain skills are better prepared to address a range of human experiences, identities, and practices. With an increased emphasis on developing and enhancing the affective domain, assessment of students’ competence in the affective domain is essential, as it provides guidance for curricular and cocurricular development. This study provides further evidence of validity for a comprehensive situational judgment test aimed at assessing the affective domain among pharmacy students. The PAD-S can be a useful instrument in the tool kit of pharmacy educators as part of a multifaceted assessment plan and offers several benefits because it is efficient, more objective, and less resource intensive than other measures, such as student reflections.

As an assessment, the internal consistency estimates (ie, reliability) of the PAD-S were generally within the acceptable range. Other situational judgment test measures have established higher internal consistency estimates; however, many of these measures were designed to assess one area or element (ie, empathy, professionalism), whereas the PAD-S includes multiple elements. Additionally, the PAD-S contains 15 items, while other situational judgment test measures use more (25 scenarios and 290 items nested within 40 scenarios). While increasing the number of items on the PAD-S would likely increase the reliability estimates, the PAD-S was designed to be completed in 20-30 minutes to avoid response burden, which was an issue noted during cognitive interviews for the longer version. Also, based on the results of person reliability and separation, potential enhancements of the PAD-S could include more challenging items to allow for assessment of all student abilities, particularly higher performing students. However, it is important to note that the purpose of the instrument is to capture key aspects of all elements of the affective domain without being time intensive, providing a comprehensive evaluation of students’ competence to determine gaps in students’ skill sets that could be addressed within the curriculum and/or the cocurriculum. As such, the level of difficulty is appropriate for this assessment because most students will likely perform at a satisfactory level, and it can still identify students who need attention and elucidate gap areas in the curriculum and cocurriculum.

The results of this study should be interpreted given a few limitations. The 15-item version included data collected from only three schools; thus, the PAD-S’s generalizability may not be of the same level for all pharmacy students. Additionally, there were variations in how the instrument was administered in the three schools, which may have affected students’ responses. For example, at one institution, it was administered alongside other survey instruments as part of annual assessment data collection processes. At another institution, the PAD-S was administered during the same time as an objective structure clinical examination (OSCE); however, the students were informed that their performance on the PAD-S would not affect their OSCE grade. Therefore, students may not have answered to the best of their ability because there were no major consequences for poor performance. However, this less stringent method may have rendered a closer reflection of how they would respond to situations or demonstrate their affective skills in a real-world situation. Another limitation could be related to the number of items.
While this study limited the PAD-S to only 15 items, more items could provide better discrimination regarding the abilities of students. However, student respondents from the previous, longer version indicated that more items led to burden; thus, it is important to balance determination of ability alongside cognitive burden. Finally, student responses to each question’s situation may not be a reflection of their own development but what they would expect others to do; this phenomenon is an inherent difficulty in measuring individual inner growth.

Despite some of its limitations, the PAD-S does have potential utility for its originally intended purpose as a formative assessment and not as a high-stakes exam. Its development involved several rounds of pilot testing, including cognitive interviewing and examining data from various versions until the current version was generated and administered. In addition, using Rasch modeling provides an approach to estimate reliability and explore internal structure validity evidence via the use of fit statistics and person-item maps that go beyond Cronbach alpha and other more traditional approaches to measuring reliability.

Based on the findings, there is no recommended scoring approach for the PAD-S, as both scoring approaches yielded fairly similar results in terms of reliability and internal structural validity. Two methods were assessed, as there was considerable debate about the ease of scoring the assessment as well as whether differentiating which of the two middle-ranked items were better or worse than the other provided meaningful information. Given the similar results, institutions may choose the scoring approach most relevant to the goals of the assessment. For example, if the goal of the assessment is to determine whether students can recognize the best and worst approach to a scenario only or to choose a simpler assessment approach, then scoring option 2 would best serve that institution. If the institution desires more discrimination, scoring option 1 (assigning scores to all four items) would be best.

Currently, the PAD-S has been incorporated in the assessment plans of one of the institutions to examine student’s affective domain skill sets during the P1 (baseline) and P3 years. Another institution has incorporated it as part of annual data collection for use in triangulation with other affective domain assessments. Future studies on the PAD-S should investigate pharmacy students’ growth with the affective domain to determine whether the instrument can measure student progress as they advance through the professional program. Additionally, it would be important to find a way to measure how often and how consistently the students demonstrate these skills and whether there are certain competency threshold levels.

CONCLUSION

A 15-item PAD-S instrument was developed to formatively measure the affective domain skills of Standard 3 and 4 for pharmacy students. Overall, the instrument performed well regardless of scoring approach. This validity evidence suggests that the PAD-S may be a useful and less resource intensive formative tool as part of a comprehensive assessment plan to identify gaps in student skill sets. Future studies involving the PAD-S should assess the measure with other samples and examine its ability to assess growth or change on the affective domain.

REFERENCES

13. Peeters MJ, Beltyukova SA, Martin BA. Educational testing and validity of conclusions in the scholarship of teaching and