Initial Validation Evidence for Clinical Case Presentations by Student Pharmacists

Jennifer S. Byrd, PharmD, BCACP, BC-ADM[1]; Michael J. Peeters, PharmD, PhD, FCCP, BCPS[2]
[1]Union University College of Pharmacy
[2]University of Toledo College of Pharmacy & Pharmaceutical Sciences

ABSTRACT

Objective: *There is a paucity of validation evidence for assessing clinical case-presentations by Doctor of Pharmacy (PharmD) students. Within Kane's Framework for Validation, evidence for inferences of scoring and generalization should be generated first. Thus, our objectives were to characterize and improve scoring, as well as build initial generalization evidence, in order to provide validation evidence for performance-based assessment of clinical case-presentations.*

Design: *Third-year PharmD students worked up patient-cases from a local hospital. Students orally presented and defended their therapeutic care-plan to pharmacist preceptors (evaluators) and fellow students. Evaluators scored each presentation using an 11-item instrument with a 6-point rating-scale. In addition, evaluators scored a global-item with a 4-point rating-scale. Rasch Measurement was used for scoring analysis, while Generalizability Theory was used for generalization analysis.*

Findings: *Thirty students each presented five cases that were evaluated by 15 preceptors using an 11-item instrument. Using Rasch Measurement, the 11-item instrument's 6-point rating-scale did not work; it only worked once collapsed to a 4-point rating-scale. This revised 11-item instrument also showed redundancy. Alternatively, the global-item performed reasonably on its own. Using multivariate Generalizability Theory, the g-coefficient (reliability) for the series of five case-presentations was 0.76 with the 11-item instrument, and 0.78 with the global-item. Reliability was largely dependent on multiple case-presentations and, to a lesser extent, the number of evaluators per case-presentation.*

Conclusions: *Our pilot results confirm that scoring should be simple (scale and instrument). More specifically, the longer 11-item instrument measured but had redundancy, whereas the single global-item provided measurement over multiple case-presentations. Further, acceptable reliability can be balanced between more/fewer case-presentations and using more/fewer evaluators.*

Keywords: case presentations, reliability, validation, generalizability theory

DESCRIPTION OF PROBLEM

Performance-based assessments are vital in pharmacy education. For this reason, well-intended educators aim for rigorous learning assessments in their courses. With a rise in this type of more-authentic performance-based assessment, interpretation and use of its scores may be based only on good intentions and untested assumptions. However, evidence for accuracy and precision in test-scores is needed to ensure educational standards have been attained by Doctor of Pharmacy (PharmD) students; especially when they are used in high-stakes decision-making.[1,2] Helping pharmacy educators assess PharmD students' readiness for advanced pharmacy practice experiences (APPEs) should be viewed as high-stakes, with consequences of delaying a student's progression in their PharmD program. One example of a performance-based assessment used in helping to assess students' preparedness for APPEs is clinical case-presentations.

There is a paucity of validation evidence for these types of assessments, especially in pharmacy education. Additionally, approaches to creating validation evidence for a performance-based assessment can be more complex than for a single written examination. For this performance-based learning assessment, we used Kane's Framework for Validation, which emphasizes validation as a process that tests a hypothesis through prioritizing and testing key assumptions with the end result being a culmination of validation evidence to support or refute the hypothesis. In this pilot, we focused on the inferences of *scoring* and *generalization*.[3] With the paucity of prior literature, we began with *scoring* evidence because this could not be assumed for clinical case-presentations. From there, we moved to initial *generalization* evidence. While more confirmatory *generalization* evidence should be sought, initial estimates could be determined from this initial administration. This report expands the limited reports of validity for test-scores from performance-based assessments, specifically clinical case-presentations in pharmacy education.

This report is the last (fourth) example in a series of articles demonstrating uses of Generalizability Theory (G-Theory) within pharmacy education. This series began with a primer on G-Theory and its use as a general methodology.[4] More specifically, the innovation of this investigation was to, within Kane's Framework for Validation, illustrate evidence for both the *scoring* and *generalization* inferences applied to pharmacy education.

Corresponding author: Michael J. Peeters, PharmD, PhD
University of Toledo College of Pharmacy & Pharmaceutical Sciences
Email: michael.peeters@utoledo.edu

DOI: https://doi.org/10.24926/iip.v12i1.2136

DESCRIPTION OF INNOVATION

Participants

In a PharmD course specifically designed for these clinical case presentations, third-year PharmD students worked up patient cases from a local hospital. Intermittently throughout the Fall semester, students orally presented five unique patient cases, including a critique/defense of those patients' therapeutic care plans, to an audience of one or two pharmacist faculty/preceptors and a small group of student peers.

Assessment

This assessment was considered part of readiness assessment before APPEs, which began the following semester. One or two faculty/preceptors (evaluators) scored presentations using an 11-item instrument with a 6-point rating-scale used previously in this course. For different students, the specific evaluator(s) differed. Items assessed included drug knowledge, disease state knowledge, patient assessment, therapeutic plan development, evidence-based recommendations, patient safety, written and oral communication, and three professionalism objectives (Supplementary File 1). This study was IRB-approved as exempt by Union University.

Statistical Analysis

This investigation was divided into two phases. One phase focused on validation evidence for the scoring inference, and the second phase focused on the generalization inference.

The first phase (scoring inference) investigated the assessment instrument itself. The Rasch Measurement Model was used for this phase via the Facets software (version 3.64.0, Winsteps.com, Beaverton OR). Linacre's recommendations for effectiveness of rating-scales featured prominently in interpretation of this analysis.[5] From Linacre's guidance, Panel A of Supplementary File 2 shows better versus worse rating-scale probability curves. These should appear as independent/prominent "peaks of hills" similar to peaks 1, 3, 5, 7, and 9, rather than the obscured "peaks" of 2, 4, 6, and 8. With a working rating-scale, meaningful summary indices could then be distilled and reported. Moreover, Wright Maps (person-item maps) could be generated for both the 11-item and the global-item instruments. Within these Wright Maps, the participants could visually be compared to the item(s), as well to the raters. Additionally, these Wright Maps could illustrate how the participants were separated into groups with the items and raters.

The second phase (generalization inference) of this investigation focused on the numbers of cases and raters. Validation evidence for this came from G-Theory analyses using the mGENOVA software (University of Iowa, Iowa City IA) for multivariate G-Theory. As an extension of Classical Test Theory, G-Theory enabled investigators to concomitantly analyze multiple error sources rather than analyzing one error source at a time as with Classical Test Theory.[4] There were four G-Theory facets in this investigation—*students*, *number of case-presentations*, *raters* (faculty/preceptors), and specific *items* that raters scored on every presentation. (Note: One or two raters independently scoring each case-presentation. No rater scored all students; the specific raters varied among case-presentations.) This G-Theory approach was the most reasonable for this performance-based assessment due to the fact that there were multiple cases, multiple items, and many raters involved.[4,6]

Multivariate G-Theory was used because the facet of items was fixed (i.e., those were the only items and so this cannot be generalized to other possible items).[4] An additional benefit is that the variance components could be explored within each of the items (assuming some items may be more statistically discriminating than others).[4] Furthermore, the global-item could be compared with the 11-item instrument. The G-Theory assessment model had *raters* nested in *cases*, crossed with *students* on a fixed set of *items* (p x (r:c)). In this assessment design, the raters were nested in cases because the raters were not the same for all students, however, the number of raters were similar for each case-presentation. Thereafter, decision-studies were performed with variance components from the G-Study.

CRITICAL ANALYSIS

During Fall of 2017, thirty-five third-year PharmD students completed five case presentations. There were 19 females and 11 males with a mean age of 26 years at the time of this course. With ratings from 15 faculty/preceptors, over 3020 data-points underlie this investigation.

Phase 1: Scoring Inference

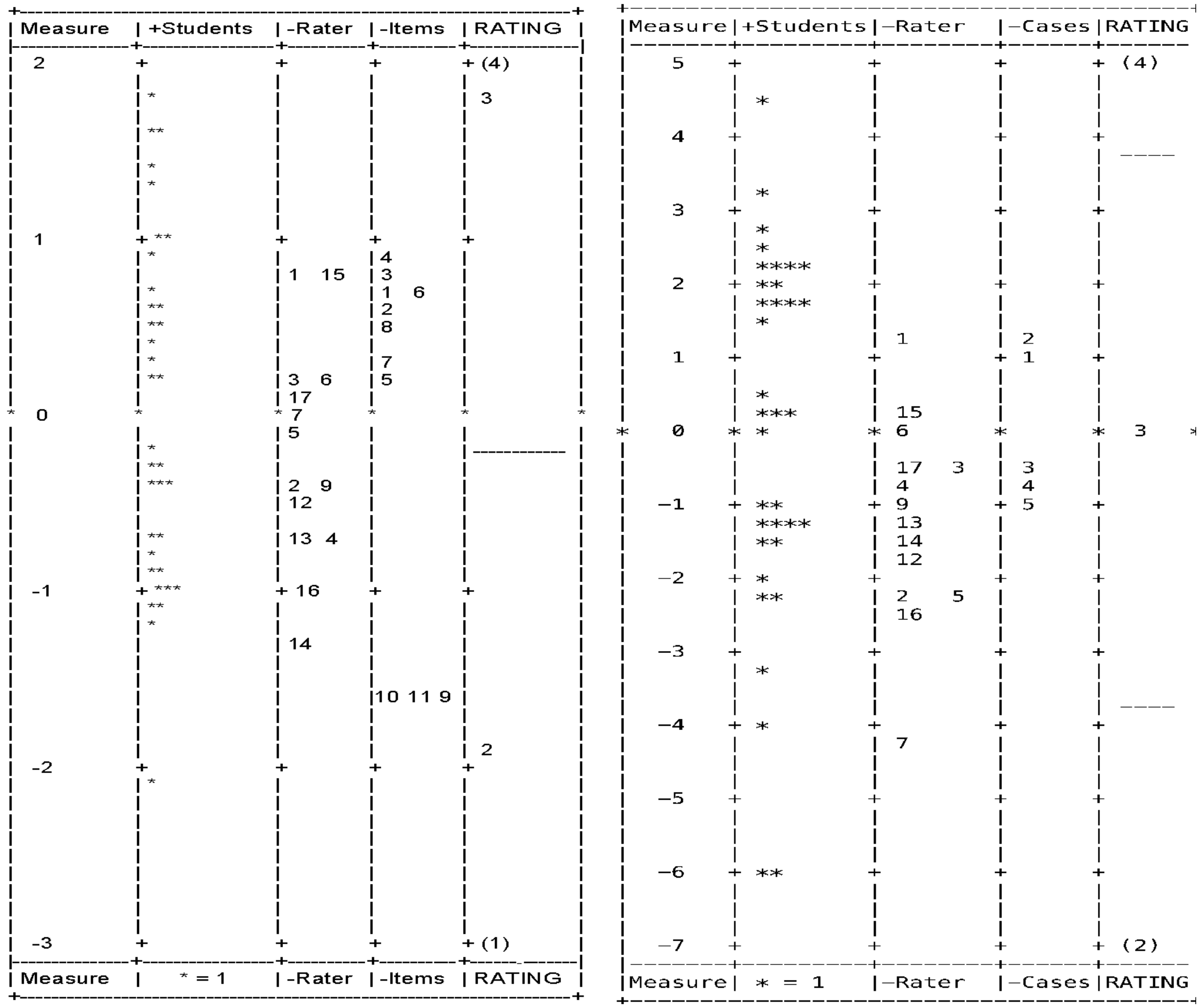
In the first phase of this investigation, we analyzed how raters actually used the rating-scale. Fifteen faculty/preceptors, seven full-time faculty and eight employed by the hospital, rated the students on their case-presentations. The 11-item instrument used a 6-point rating-scale that included a label for each rating-scale category (Supplementary File 1). This rating-scale did not appear to work, as shown in panel B of Supplementary File 2, with overlapping non-peaks for some rating-scale categories. However, the rating-scale of the 11-item instrument worked when the 6-point rating-scale was collapsed into a 4-point rating-scale. This can be seen in panel C of Supplementary File 2, with independent peaks for all categories of the 4-point rating-scale.

With this working 4-point rating-scale, the 11-item instrument was then compared to the global-item. For the 11-item instrument, overall measurement indices were a separation of 8.62 and reliability of 0.99 among these participants. For the global-item, measurement indices were a separation of 2.57 and reliability of 0.87 among these participants. Wright Maps for these analyses are in Panels A & B of Figure 1. In both panels, the first column of measure is the common scale for the other columns (i.e., to compare distributions across columns). Across columns in both panels, the distribution of students,

DOI: https://doi.org/10.24926/iip.v12i1.2136

raters, cases, and most items aligned with distributions in the other columns; most of these appeared adequate to measure (not too easy or too hard). Measures of students' case-presenting ability, from less (bottom) to more (top) able, were well distributed in both panels. Similarly, raters were easier (bottom) to harder (top) in both panels. In Panel A of Figure 1, the items of the 11-item instrument were sequenced from easier (bottom) to harder (top); wherein similar horizontally-positioned items shared similar measures (and so may be redundant). Items 9-11 were lower than all students and so were too easy for everyone (i.e., can be removed as not helping to measure anything). An item-difficulty column was not needed because analysis had only one global-item; instead, a case-presentation-difficulty column replaced item-difficulty in Panel B of Figure 1. Shown in Panel B for the global-item, cases were ordered from easiest (bottom) to hardest (top). The case-presentations were roughly sequenced in order, with most students finding case-presentations 1 & 2 most difficult, while case-presentations 3-5 were easier in this course. That is, as students became more proficient with case-presentations, this activity became easier for them.

Figure 1. Rasch Measurement Wright Maps (person-item) for an instrument to score PharmD case presentations



Panel A. 11-item instrument

```
+---------------------------------------------------------------+
| Measure      | +Students     | -Rater  | -Items  | RATING     |
|--------------+---------------+---------+---------+------------|
|  2           +               +         +         + (4)        |
|              |               |         |         |            |
|              | *             |         |         | 3          |
|              |               |         |         |            |
|              | **            |         |         |            |
|              |               |         |         |            |
|              | *             |         |         |            |
|              | *             |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|  1           + **            +         +         +            |
|              | *             |         | 4       |            |
|              |               | 1   15  | 3       |            |
|              | *             |         | 1   6   |            |
|              | **            |         | 2       |            |
|              | **            |         | 8       |            |
|              | *             |         |         |            |
|              | *             |         | 7       |            |
|              | **            | 3   6   | 5       |            |
|              |               | 17      |         |            |
*  0           *               * 7       *         *            *
|              |               | 5       |         |            |
|              | *             |         |         | ---------- |
|              | **            |         |         |            |
|              | ***           | 2   9   |         |            |
|              |               | 12      |         |            |
|              |               |         |         |            |
|              | **            | 13  4   |         |            |
|              | *             |         |         |            |
|              | **            |         |         |            |
|  -1          + ***           + 16      +         +            |
|              | **            |         |         |            |
|              | *             |         |         |            |
|              |               | 14      |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |10 11 9  |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         | 2          |
|  -2          +               +         +         +            |
|              | *             |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|              |               |         |         |            |
|  -3          +               +         +         + (1)        |
|--------------+---------------+---------+---------+------------|
| Measure      |     * = 1     | -Rater  | -Items  | RATING     |
+---------------------------------------------------------------+
```

Panel B. Global-item

```
+------------------------------------------+
|Measure|+Students|-Rater   |-Cases|RATING
|-------+---------+---------+------+-------
|   5   +         +         +      + (4)
|       |         |         |      |
|       | *       |         |      |
|       |         |         |      |
|   4   +         +         +      +
|       |         |         |      | ----
|       |         |         |      |
|       | *       |         |      |
|   3   +         +         +      +
|       | *       |         |      |
|       | *       |         |      |
|       | ****    |         |      |
|   2   + **      +         +      +
|       | ****    |         |      |
|       | *       |         |      |
|       |         | 1       | 2    |
|   1   +         +         + 1    +
|       |         |         |      |
|       | *       |         |      |
|       | ***     | 15      |      |
*   0   * *       * 6       *      *  3   *
|       |         |         |      |
|       |         | 17   3  | 3    |
|       |         | 4       | 4    |
|  -1   + **      + 9       + 5    +
|       | ****    | 13      |      |
|       | **      | 14      |      |
|       |         | 12      |      |
|  -2   + *       +         +      +
|       | **      | 2    5  |      |
|       |         | 16      |      |
|       |         |         |      |
|  -3   +         +         +      +
|       | *       |         |      |
|       |         |         |      |
|       |         |         |      | ----
|  -4   + *       +         +      +
|       |         | 7       |      |
|       |         |         |      |
|       |         |         |      |
|  -5   +         +         +      +
|       |         |         |      |
|       |         |         |      |
|       |         |         |      |
|  -6   + **      +         +      +
|       |         |         |      |
|       |         |         |      |
|       |         |         |      |
|  -7   +         +         +      + (2)
|-------+---------+---------+------+-------
|Measure| * = 1   |-Rater   |-Cases|RATING
+------------------------------------------+
```

Measure = in logarithm-odds units, this is a common scale for the next four columns
Students = distribution of students (higher = more able)
Raters = distribution of raters (higher = harder)
Items = distribution of items (higher = harder)
Cases = distribution of cases (higher = harder)
Rating = distribution of ratings (low to high)

DOI: https://doi.org/10.24926/iip.v12i1.2136

Phase 2: Generalization Inference

In the second phase of this investigation, we analyzed how many cases would be needed to achieve acceptable reliability of scores from this performance-based assessment. Scores from the 6-point rating-scale were revised to a 4-point scale before this G-Theory phase. For this learning assessment, the g-coefficient (reliability) for the series of five case presentations was 0.76 using the 11-item instrument and 0.78 using the global-item.

Of the 11-items, three items had no variation; all students received the same score on all case-presentations. Due to lack of variation, those three professionalism items were removed from this G-Theory analysis (These are noted in Supplementary File 3). The variance components are in Supplementary File 3 for each (fixed) item in the instrument as well as the global-item. Among the remaining eight items, there were notable discrepancies in variances from different facets. For all items, the *rater* facet showed little variance, while *student x case* showed much more (of note, this *student x case* interaction is more generally termed case-specificity).

Decision-Studies. Table 1 shows the estimated g-coefficients (reliability) for one-rater in Panel A and two-raters in Panel B. Within these panels, the effect of increasing cases can be seen for individual items, for the cumulative 8-items (from the 11-item instrument), and for the global-item.

Table 1. Estimated G-coefficients for a number of specific items used to evaluate PharmD case-presentations

Panel A. One Rater

	Number of Case-Presentations (for Each Item)									
Item Number	1	2	3	4	5	6	7	8	9	10
Item 1	0.32	0.49	0.59	0.66	0.70	0.74	0.77	0.79	**0.81**	**0.83**
Item 2	0.37	0.54	0.64	0.70	0.75	0.78	**0.81**	**0.83**	**0.84**	**0.86**
Item 3	0.16	0.27	0.36	0.43	0.49	0.53	0.57	0.60	0.63	0.65
Item 4	0.21	0.28	0.32	0.34	0.35	0.36	0.36	0.37	0.37	0.38
Item 5	0.18	0.30	0.39	0.47	0.52	0.57	0.60	0.64	0.66	0.69
Item 6	0.29	0.37	0.40	0.42	0.43	0.44	0.45	0.45	0.46	0.46
Item 7	0.25	0.40	0.50	0.57	0.62	0.67	0.70	0.73	0.75	0.77
Item 8	0.32	0.49	0.59	0.65	0.70	0.74	0.77	0.79	**0.81**	**0.83**
Total Score Items 1-8	0.36	0.54	0.64	0.70	0.75	0.78	**0.81**	**0.83**	**0.84**	**0.86**
Global Item	0.35	0.52	0.62	0.68	0.73	0.76	0.79	**0.81**	**0.83**	**0.84**

Panel B. Two Raters

	Number of Case-Presentations (for Each Item)									
Item Number	1	2	3	4	5	6	7	8	9	10
Item 1	0.39	0.56	0.66	0.72	0.76	0.79	**0.82**	**0.84**	**0.85**	**0.87**
Item 2	0.45	0.62	0.71	0.77	**0.80**	**0.83**	**0.85**	**0.87**	**0.88**	**0.89**
Item 3	0.20	0.34	0.43	0.51	0.56	0.61	0.64	0.67	0.70	0.72
Item 4	0.28	0.34	0.36	0.37	0.38	0.38	0.39	0.39	0.39	0.40
Item 5	0.24	0.39	0.49	0.56	0.61	0.66	0.69	0.72	0.74	0.76
Item 6	0.45	0.54	0.57	0.59	0.60	0.61	0.62	0.62	0.63	0.63
Item 7	0.32	0.49	0.59	0.66	0.70	0.74	0.77	0.79	**0.81**	**0.83**
Item 8	0.40	0.58	0.67	0.73	0.77	**0.80**	**0.83**	**0.84**	**0.86**	**0.87**
Total Score Items 1-8	0.38	0.56	0.66	0.72	0.76	0.79	**0.82**	**0.84**	**0.85**	**0.86**
Global Item	0.42	0.59	0.68	0.74	0.78	**0.81**	**0.83**	**0.85**	**0.87**	**0.88**

Note: Grey, bolded cells are at or beyond a threshold of 0.8[2]

KEY ISSUES

Our findings confirmed that, with this evaluation tool, scoring should be simple in terms of both the rating-scale and the entire instrument.[7-9] Regarding the rating-scale, all points on the collapsed 4-point rating-scale were used. While raters requested more points on the rating-scale (i.e., six categories), the 6-point rating-scale did not work (Shown in Panel B of Supplementary File 2). This finding is aligned with recommendations for rater judgment to use a 4-point rating-scale.[8] Moreover, this performance-based assessment instrument evolved over multiple course iterations (over multiple years) to include 11-items. Raters affirmed that each of the 11 items was an important consideration for clinical case-presentation by PharmD students before progressing to their APPEs; though, this instrument showed redundancy. The global-item appeared to provide measurement potential from Rasch Analysis, and its measurement potential was further shown with the G-Theory evidence.

Our multivariate G-Theory Analysis allowed comparison for trade-offs from using one or two raters, and the 11-item instrument or global-item. The global-item statistically discriminated similarly to the entire 11-item instrument. As expected, two raters were somewhat more reliable than one rater. Using the 11-item instrument scored by either one or two raters, an estimated seven case-presentations would be needed for adequate reliability. However using only the global-item, an estimated six case-presentations scored by two raters (i.e., twelve raters) or eight case-presentations scored by one rater (i.e., 8 raters) could be adequately reliable. Thus, having more case-presentations seems preferred.

The global-item appeared comparable to the much longer 11-item analytic rubric. As with the rating-scale, simplicity appears to prevail here as well. Although, one notable advantage of an analytic rubric is that it can provide more specific feedback to a student for the grade they earned. Thus, sharing this completed analytic rubric with students has been a helpful priority for feedback at Union University College of Pharmacy. A 'mixed approach' to rubric development may help;[10] providing specific analytic-rubric feedback to students, as well as a global-item that makes it quicker for the course coordinator to calculate course grades.

It is noteworthy that the *student* facet did not contain all the variance (as some educators might assume). While less than one-third of total variance was due to students, there were other sizable variance sources. The *cases* facet showed some variance, and it was notable in the Rasch Wright Map (Figure 1, Panel B) that cases were roughly numbered sequentially in difficulty (with 1 & 2 more difficult than 3-5). Case-specificity, an interaction of the *student* and *case* facets in this performance-based assessment, is a common scourge of learning assessments.[6,7] It is routine to use multiple tasks, such as multiple case-presentations, to overcome reliability limitations from case-specificity. This is demonstrated in Table 1. Additionally, minimal variance came from the *rater* facet. While using these pharmacy practice experts as raters often relies on "subjective" expert judgments, an acceptable, rigorous, objective reliability can be attained, similar to other learning assessment methods (such as multiple-choice question testing), provided that adequate sampling of tasks (case-presentations) is done. Additionally, using multiple case-presentations (and so multiple raters) also has the advantage that no single rater makes the difference between a student doing poorly or doing well on the entire series of case-presentations. Thus, these are multiple reasons for using multiple tasks in a performance-based assessment.

A major limitation is noteworthy. While broad findings will generalize to other programs, the specific numbers for items, raters, and case-presentations may differ, as these are context-dependent and sample-dependent to this one institution. While educators elsewhere can gather suggestions from this performance-based assessment, each PharmD program should, ultimately, evaluate their own learning assessments - especially for high-stakes use.[2,3] Validation evidence is needed locally, for a particular cohort of students' scores and in the specific educational context. Furthermore, the global-item was scored only after raters completed the 11-item analytic rubric. Therefore, this was not exclusively a one-item (holistic) rubric; instead, it was one global-item after raters considered 11 criteria. Thus, a mixed approach to the revised rubric seems especially prudent given the raters' use of this case-presentation instrument.[10]

At Union University, the most desired condition was a reduction in preceptor resources (i.e., time spent scoring the presentations followed by fewer raters). The number of clinical case-presentations by students could practically be increased. In fact, over a series of two academic courses, students complete 9 case-presentations; and following from our D-Studies (with one or two raters and 11-item or global-item), all options appear sufficiently reliable for high-stake testing. Thus, the global-item (within a mixed rubric) scored by one rater seems best in this particular context. It seems easiest for faculty/preceptors, and easier for the course coordinator to collate into grades. Moreover, it may become even better, if the original 6-point, 11-item instrument is revised into a mixed-rubric[10] with only a 4-point rating-scale; this would require fewer case-presentations.

NEXT STEPS

As evaluation of this pilot study demonstrated, validation evidence should support changes (including unexpected outcomes), necessitating future iterative evaluation of this performance-based assessment. Validation evidence for inferences of *extrapolation* and *implications* (within Kane's Framework for Validation) have not yet been systematically explored for scores from this learning assessment. Future investigations of these would be advised, especially for decision-rules that are evidence for the *implications* inference.

DOI: https://doi.org/10.24926/iip.v12i1.2136

CONCLUSION

Test validation is specific to the educational context and is specific to test-scores from each learning assessment at every specific college/school of pharmacy. Within test validation, evidence for scoring and generalization inferences are vital. Scoring is not always straightforward; a 4-point scale vastly improved scores for measurement with this instrument, and measurement parameters using only the global-item also looked promising. Following the improved scoring, various one-rater and two-rater scenarios, along with 11-item and global-item scenarios, were explored. In the end, validation evidence for the scoring and generalization inferences were empirically evaluated.

Conflicts of Interest: None
Funding/support: None

REFERENCES

1. Accreditation Council for Pharmacy Education. Accreditation Standards and Key Elements for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree ("Standards 2016"). Published February 2015. Available at: https://www.acpe-accredit.org/pdf/Standards2016FINAL.pdf. Accessed 3 Jan 2021.
2. Peeters MJ, Cor MK. Guidance for high-stakes testing within pharmacy education. *Curr Pharm Teach Learn*. 2020; 12(1): 1-4. doi: 10.1016/j.cptl.2019.10.001
3. Peeters MJ, Martin BA. Validation of learning assessments: a primer. *Curr Pharm Teach Learn*. 2017 Sep 1;9(5):925-933. doi: 10.1016/j.cptl.2017.06.001
4. Peeters MJ. Moving beyond Cronbach's alpha and inter-rater reliability: A primer on Generalizability Theory for pharmacy education. *Innov Pharm*. 2021; 12(1):Article 14.
5. Linacre JM. Optimizing rating scale category effectiveness. *J Appl Meas*. 2002; 3(1):85-106.
6. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales*. 5th ed. New York, NY: Oxford University Press; 2015.
7. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996; 1(1):41-67. doi: 10.1007/BF00596229
8. van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ*. 2005; 39(3):309-317. doi: 10.1111/j.1365-2929.2005.02094.x
9. Peeters MJ. Measuring rater judgment within learning assessments—Part 1: Why the number of categories matters in a rating scale. *Curr Pharm Teach Learn*. 2015;7(5):656-661. doi: 10.1016/j.cptl.2015.06.015
10. Peeters MJ. Measuring rater judgment within learning assessments—Part 2: A mixed approach to creating rubrics. *Curr Pharm Teach Learn.* 2015; 7(5):662-668. doi: 10.1016/j.cptl.2015.06.022

DOI: https://doi.org/10.24926/iip.v12i1.2136