Validation of Authentic Performance Assessment: A Process Suited for Rasch Modeling

Terrence R. Jackson, JoLaine R. Draugalis, Marion K. Slack and Woodie M. Zachry
College of Pharmacy, The University of Arizona, PO Box 210207, Tucson AZ 85721-0207

Jerome D’Agostino
College of Education, The University of Arizona, PO Box 210207, Tucson AZ 85721-0207

An evaluation was conducted in order to develop, test, and validate a method for assessing critical thinking learning outcomes as related to scientific literature evaluation in a PharmD program. A 25-item authentic assessment was created using recommended guidelines and Bloom’s hierarchical taxonomy. Construct and content validity were supported using classical test theory and Rasch modeling. A primer on using Rasch analysis is provided in conjunction with the specific results. Rasch analysis demonstrated good INFIT and OUTFIT statistics, with item difficulty ranging from -1.90 to 2.46 logits for 23 items. The KR20 for the assessment was 0.41 and item point-biserial correlations ranged from -0.05 to 0.25. Rasch modeling techniques provided information to evaluate individual item contributions not discerned by standard techniques such as KR20 or point-biserial correlations. This research supports the practical application of this type of authentic assessment and provides evidence for construct and content validity in measuring student performance.

INTRODUCTION
Continuing improvements in education have focused on enhancing the understanding and the development of skill mastery in students. This is especially true in the education of health professionals such as pharmacists, physicians, dentists, or nurses. It is essential that the student is able to perform the tasks of his or her profession accurately and consistently as a result of their education. This outcomes or abilities-based education, represents a shift in basic thinking about assessment and places emphasis on performance and standards of competency (i.e., mastery)(1). Tools are needed to effectively facilitate evaluation and self-learning of critical competencies(2). As Wiggins stated in an interview “First, we need to provide teachers with models and criteria for what good performance assessment looks like”(3). The task then is developing and applying the best assessment tool to fit the desired competency(4). Hence, performance assessment is an integral component of abilities-based education where an emphasis lies in what the students are able to do as a result of their instruction(5). There are two basic design criteria of any performance __________

1Corresponding author address: College of Pharmacy, (M/C 871), University of Illinois at Chicago, 833 South Wood Street, Chicago IL 60612-7230.
assessment. The first is that it is meaningful and authentic, the second is one of validation(6).

While a concise definition of authentic assessment eludes consensus, it is purported to be a method to measure higher-order knowledge and skills in a setting that approximates the “real world”(7). Authentic assessments representative of the types of situations encountered by pharmacy students (observed simulated clinical examinations, assessment scenarios, and portfolios) have been useful in facilitating student learning(8-10). Additionally, authentic performance assessment provides a student with quality information they can use in order to improve their skills and allows further learning from the process of the assessment itself(11,12).

Validity must be thought of as an inductive summary of assessment score interpretation and its subsequent use. Validity issues must be considered part of the design template when developing an authentic assessment instrument(6). The process of validation may then lead to revisions of the authentic assessment where support for valid inference is weak. Mehrens reported that the practices of validation for authentic performance assessments were few(13). Unfortunately, this finding continues to be true.

Construct validation for any assessment at a minimum should include the following characteristics: (i) convergence; (ii) discrimination; (iii) sampling adequacy; and (iv) reliability. Convergence in an assessment is supported by evidence suggesting all the items converge to support a single domain, (i.e., a unidimensional construct). Discriminability is supported by providing evidence that allows for items that are unrelated to the construct to be determined (i.e., the assessment should not be affected by construct-irrelevant variance). This construct-irrelevant variance, the amount the authentic assessment is affected by elements outside the construct being evaluated, is a major threat that undermines validity(4,14).

A major threat to sampling adequacy is construct under-representation, (i.e., the inadequate ability to describe the important aspects of the construct). Item difficulty that is hierarchically ordered provides for the detection of construct under-representation(15). Additionally, having an assessment that is relevant to what has been explicitly taught and where deviations from performance can be detected is essential. That is, appropriate task processes that permit differentiation of student ability is important to include in domain coverage(16,17). To evaluate construct relevance, individuals who are experts in subject matter should review and critique the tasks the students are expected to perform(16). Another strategy for improving domain coverage is to use structured exercises (e.g., authentic assessment) that contain performance tasks providing information about knowledge and skills related to the performance domain.

Popham states “Test reliability is a necessary but insufficient, condition for valid score-based inferences”(18). Specifically, reliability refers to the quantity of error variance. There are several specific estimates of reliability to choose from (e.g., test-retest, generalizability theory, internal consistency). The reliability estimate chosen for use depends on the error producing factors that one seeks to identify. If error factors associated with the use of different items is of interest, then estimating score reliability is accomplished by evaluating the inner structure of an assessment instrument for unidimensionality and for item interrelationships(14).

An assessment design validation process must be chosen to address validity and reliability concerns. Two approaches are commonly used, classical test theory (CTT) and item response theory (IRT). It has been said that increased application of IRT and the movement towards authentic assessment are two of the most important developments in educational measurement(19). It should be noted that while Rasch modeling historically has been referred to as a member of the IRT family, the rapid growth in the use of Rasch modeling has led to the development of Rasch modeling as a field of measurement in its own right. To distinguish it from other forms of IRT, the term “Rasch” will be used in the remainder of this paper.

The primary differentiation between CTT and Rasch modeling is that the former is described as “test based” and the latter described as “item based”(20). Item difficulty (percent correct) and item discrimination (point-biserial correlations) in classical testing, are both dependent upon the characteristics of the sample of students that have been tested(20). Historically, the point-biserial correlation has been an accepted method for evaluating item reliability. Pray and Popovich used a threshold of 0.20 or higher as an indicator for item acceptability in screening 300 test items tested on more than 880 students for inclusion into a final competency assessment instrument(21). As well, at Wayne State University a threshold point-biserial value of 0.20 was used to screen 286 items field-tested on 196 students for inclusion into a comprehensive examination(22). Both of these studies demonstrate how large numbers of items and students were used to obtain reliable statistical analyses.

There are two clear limitations to the use of the point-biserial. First, the requirement for large numbers of students and items limits its usefulness for typical class sizes and examinations. Secondly, there are conventions, but no clear interpretation of what constitutes an acceptable value when evaluating an item. Because the point-biserial is greatly influenced by the score distribution of the sample, it approaches zero as the variance in scores decreases(23). Thus, if the variance decreases enough, the point-biserial may fall below the conventional threshold of 0.20 regardless of the item quality. This information may be misleading when attempting to evaluate an individual items contribution to an assessment.

The purpose of this study was to develop an authentic assessment and apply Rasch analysis and CTT to validate its performance as a method for assessing levels of learning outcomes using, as a guide, material specific to domains relating to professional development, specifically professional/scientific literature evaluation (i.e. critical analysis) skills that include study design topics, statistical techniques, safety, efficacy, and cost-effectiveness in an entry-level PharmD program.

**Primer on Rasch Modeling**

The Rasch model is a mathematical model that describes the relationship between the probability of correctly answering an item (P) and the difference between the person’s ability (B) and the item’s difficulty (D), expressed by the formula: \( P = \frac{\exp(B-D)}{(1 + \exp(B-D))} \)(24). The dichotomous Rasch model takes into account only differences in item difficulty to provide a measure of student ability(25). Rasch analysis compares each person and item in terms of fitting a unidimensional model (i.e., a single construct underlying the items). The Rasch model has two requirements, Unidimensionality and local independence. First, items must work together to define a single unidimensional construct (unidimensionality) and demonstrate hierarchical
structure such that the items form a continuum in which they are easy at one end and difficult at the other. Bejar stated “Unidimensionality does not imply that performance on items is due to a single cognitive process”(26). To be useful the items need only function in unison, measures need only to behave as though they are unidimensional(27,28). For example, the assessment of skills related to critical evaluation of professional/scientific literature may require the student to consider topics related to study design, statistical techniques, efficacy, safety, and cost effectiveness. While individually these aspects may function independently, together these components may function unidimensionally defining the underlying construct. The capability of evaluating the contributions of content areas that work together in the real-world environment in a single assessment is one of the major strengths of Rasch modeling. It is this feature of Rasch modeling that permits computer adaptive testing of many board licensure examinations including those used in pharmacy. The second requirement (local independence) is that the probability of the student correctly responding to an item does not depend on the other items in which the assessment or the order in which the items appear.

An advantage to using Rasch analysis is that relatively few subjects are needed [50] to sufficiently obtain useful and reasonable estimates. For well designed pilot studies, 30 subjects may be sufficient(29). This is in contrast to methods like the KR20 and point-biserial correlation which are dependent upon the number of items in the test and the distribution of students’ scores(30).

The production of scale scores using Rasch analysis may be beyond what some educators might use in the classroom. However, Popham states that “educators will be forced to understand something about such score-interpretation methods”(18). A brief description of Rasch concepts and associated terminology used in the evaluation of performance and validity are provided to familiarize the reader with aspects of its use. Unidimensionality and local independence in Rasch analysis are assessed by item FIT statistics. The degree of agreement between the pattern of observed responses and the modeled expectations is described using FIT (INFIT and OUTFIT) statistics. The requirement of unidimensionality and local independence are met when the data fit the model and reliability of item placement is established (i.e., the criteria set for Fit statistics and reliability indices are met). These statistics provide empirical evidence to detect: (i) when an item is not part of the same dimension being measured; (ii) the item is not understood; and (iii) when it is likely that the response is a guess or a student possesses special knowledge.

FIT statistics are reported in two forms, INFIT and OUTFIT. INFIT is a weighted statistic that is sensitive to unexpected behavior that affects responses to items near the person’s ability level. An INFIT statistic indicates the degree to which the observations for a particular item meets the model expectations. For example, an item with an INFIT Zstd of greater than 2.0 (greater than two standard deviations) would suggest that the item may be measuring something different than intended (e.g., a different construct). OUTFIT, an unweighted statistic, is an outlier-sensitive fit statistic that indicates whether unexpected responses or outliers are found based on the person’s ability. For example, an OUTFIT Zstd of greater than 2.0 would suggest some guessing or special knowledge on the students part thus enabling a student of lesser ability to correctly respond to an item of greater difficulty. INFIT and OUTFIT are both reported in two forms, unstandardized as mean square (MNSQ) and the standardized MNSQ (Zstd). The MNSQ indicates the amount of randomness in the response pattern when the data fit the model. It shows the magnitude of the discrepancy between the observed and expected responses. The standardized mean square statistic approximates the t-statistic with a mean 0.0 and variance 1.0 distribution. This Zstd shows the statistical probability of the discrepancy. Both high and low INFIT and OUTFIT statistics are evaluated.

Item difficulty is described on a continuum from less difficult to more difficult and is calibrated in logits. A logit is a unit of measurement used in Rasch analysis for calibrating items and measuring persons, based on the natural logarithmic odds of the probability of a response. The validity of interpreting items in terms of difficulty are verified by examining whether all the items are hierarchically ordered to define the construct (28). Item difficulty is calibrated and student ability is measured. In the Rasch model, a person’s ability is defined as the log odds of answering correctly, items of “average” difficulty on the same scale. The average item logit is centered to zero to establish this scaling. For example, a person with an ability measure of + 1.0 logits would have a 50 percent probability of correctly responding to an item of +1.0 logit difficulty.

An item distribution map is constructed to show the distribution of the persons and items on the same measurement scale (see Figure 1). The scale measuring the construct is laid out vertically with the most able persons and most difficult items at the top. The left-hand column locates the person ability as measured in logits. The right-hand column locates the item difficulty calibrations reported in logits. This map enables the researcher to visually observe the function of items and ability measures of students as a whole. Additionally, it provides the researcher a visual inspection of the spread of items along the scale. A separation distance of 0.15 logits or more between items demonstrates that items are distinct from each other(32). It should be noted that this is only a crude estimate as a more precise value would be sample specific and depend on the standard error. The greater the separation, the more distinct measurement strata that are identified. However, the separation between items should not be so large that gaps in measurement occur. For example, visible gaps (i.e, > 0.30 logits) in the map indicate regions of the assessment where the creation of additional items within that difficulty range might improve the measurement of the construct.

Rasch modeling conveys information about student ability in reference to defined traits(33). It allows the determination of student performance relative to the difficulty of the item because the item and ability statistics are reported on the same scale in Rasch modeling. This allows the researcher to know how an item performs in relation to the student’s ability (i.e., which items challenge the student). Another strength of Rasch model analyses is that it especially useful in helping to clarify
our understanding of measurement with multiple-choice items(23). Wang, states that “fitting the Rasch model is one of the best criteria for constructing multiple-choice items”(34).

METHODS

This study consisted of three sections. First, the authentic assessment was developed. Then validation was conducted using both classical test theory and Rasch modeling. Finally, feedback was gathered pertaining to usefulness of the authentic assessment

Subjects: The subjects used in this study were students enrolled in their second (year-2) and third (year-3) professional years in the doctor of pharmacy degree program at the University of Arizona during the academic year 2000-2001.

These students were selected because they have completed or were currently enrolled in courses that contribute to the content of the abilities within the domain being authentically assessed. Participation in this study was voluntary and the research was approved by the human subjects committee.

Development of the Authentic Assessment: Several principles from the American Association for Higher Education’s (AAHE) principles for good practice were used to develop the authentic assessment instrument(35). Development of the authentic assessment methodology occurred in three steps. The first step identified the competencies to be evaluated for use in the authentic assessment. In the second step, the appropriate authentic assessment method was selected. Finally, the items for the authentic assessment were developed.

The AAHE principle “The assessment of student learning begins with educational values” was used in the first step. This principle was met by incorporating the American Council on Pharmaceutical Education and The University of Arizona College of Pharmacy institutional guidelines(36). Content was developed based on these guidelines. The authentic assessment was constructed in order to address skills associated with critical thinking, including the abilities: evaluate professional/scientific information and literature in a critical, scientific, and effective manner; and evaluate the safety, efficacy, and cost effectiveness of pharmaceutical products and services. Students at the University of Arizona College of Pharmacy are exposed to this material in classes during their second and third professional years. Additionally, the student applies this material while on rotations during the final year of the program.

Step two involved choosing an appropriate type of authentic assessment. Authentic assessment of students’ critical evaluation of professional literature skills would be based on “real world” tasks required in clinical practice(37). The critical analysis of a journal article, was deemed appropriate as these skills are also required in professional practice(10). The “real world” reading task in this study was the evaluation of the journal article “Effect of an automated, nursing unit-based drug-dispensing device on medication errors” by Borel and Rascati(38). This journal article was chosen because it represents a scientific paper that includes methodological issues and material that a pharmacist needs to be proficient in evaluating. It is necessary to note however that the specific content reflected in the Borel and Rascati article is not emphasized in the curriculum. The intent of using this article was to allow the evaluation of performance and ability related to the competencies selected in a manner that did not simultaneously evaluate skills related to disease management. If an article relating to an area of expected subject content mastery (e.g., asthma) had been chosen, it may have been difficult to differentiate that ability (asthma management) from performance related to the competencies intended for authentic assessment (i.e., critical analysis skills). The AAHE principle applicable here states “Assessment makes a difference when it begins with issues of use and illuminates questions that people really care about.” Well performed assessment necessitates that there be a buy-in from the students where there is credibility in its application. This credibility may be seen when useful assessment data are relevant.

Once the authentic assessment content was determined, item development followed using Bloom’s taxonomy. The use of Bloom’s taxonomy has provided a useful template for constructing an assessment with varying levels of difficulty including items that would measure higher order thinking(39,40).
4Sultana, Q., Klecker, B.M., “Evaluation of first-year teachers’ lesson objectives beyond the critical values were then evaluated further for their predicted by the Rasch model that estimates more variation modeling. Items with OUTFIT statistics (a pattern of response was also assessed by INFIT statistics obtained from Rasch representation of the authentic assessment items was also supported using Marzano’s New Taxonomy(41).

Fig. 2. Authentic assessment evaluation indicators.

It should be noted that while items were developed ranging from knowledge through evaluation classifications, that the authentic assessment represents an application of the content areas and supports the higher order processes. This higher order representation of the authentic assessment items was also supported using Marzano’s New Taxonomy(41).

Evaluation of Validity Evidence: Validation (i.e., determining whether the assessment measures what we think it does) was a primary function of this study. Both construct and content validity were evaluated in this study. These types of validity were used as a framework in which to dissect and evaluate the intent of the authentic assessment measurements. The lack of a valid measure of competency relating to learning outcome domains for doctor of pharmacy students prohibited the measurement of criterion-referenced validity.

Six parameters were used to provide evidence of construct validity. CTT provides the KR20 and point-biserial correlation coefficient and the Rasch model generates separation reliability, INFIT and OUTFIT statistics, and the item distribution map (evaluated in that order). The KR20-if-item deleted was calculated to assess what the KR20 would be if that item was deleted from the assessment. The criteria for evaluating items are shown in Figure 2.

Item separation reliability and INFIT statistics were evaluated first to verify unidimensionality and local independence. Construct irrelevance, i.e., the item does not measure the intended construct, was also assessed by INFIT statistics obtained from Rasch modeling. Items with OUTFIT statistics (a pattern of response predicted by the Rasch model that estimates more variation than expected), KR20, and point-biserial correlations beyond the critical values were then evaluated further for their contribution to the overall authentic assessment and a determination was made to retain or delete the item. When an item exhibited borderline misfit statistics the item’s difficulty (logit value) was evaluated for its placement relative to other items. If there were no other items that permitted measurement at that logit value, an argument was made to retain the item. Construct underrepresentation was then evaluated by examining the hierarchical placement of items in the item distribution map.

Initial content validation, a judgment process that is typically done at some level with most assessments (and most often stops at this point), was supported by having a panel of five individuals familiar with the research content of this study review the items of the authentic assessment. These individuals (all pharmacists), three PhDs, one MS, and one PharmD, were given a list of the items and the article and asked to evaluate the intent of the authentic assessment to measure the ability of the students relating to the competency being evaluated. This judgment was subjective in nature. The judges agreed that the competency evaluated was supported by all of items. The judges reached their decisions independently. Additionally, because the authentic component of the assessment was a published study, the authentic assessment as a whole was deemed representative of the content of the competency.

Student Feedback: Students rated the usefulness of the authentic assessment by responding to the following questions: 1) How useful was the Authentic Assessment in helping you assess your learning needs?, 2) How useful would this type of assessment be in other areas of the curriculum?, 3) I would recommend this type of assessment be performed before beginning rotations. The first two questions were rated using a three-point rating scale ranging from not at all useful to very useful. The last question was rated using a six point scale ranging from strongly agree to strongly disagree.

STATISTICAL ANALYSIS
Data (containing dichotomously scored results of the authentic assessment) were entered into a data file with MS-DOS which were input into Winsteps version 3.07 in order to calculate statistics for the Rasch model(42). The output generated from Winsteps included point-biserial correlations, separation reliability (separation index and item reliability), item INFIT and OUTFIT statistics, and the item difficulty map. SPSS statistical analysis system version 9.0 for windows was used to calculate the KR20.

RESULTS
The authentic assessment was administered to 80 volunteers of 109 students available at the time of the study representing a sampling of 73.4 percent of the students. The sample consisted of 44 out of 52 second-year (class of 2002) and 36 out of 57 third-year (class of 2003) students. The mean number of correct items for the second-year class was 12.36, SD = 2.54 (49.4 percent) and for the third-year class was 10.28, SD = 2.65 (41.1 percent). All students completed the assessment within 50 minutes.
The summary statistics for the Rasch model are presented in Table I. The reliability separation index for the 25 item assessment was 4.20 and item reliability was 0.95 indicating items create a variable that is well defined and that the reliability of item placement along the scale is good. Evaluation of INFIT statistics for the items in the assessment show that MNSQ values were less than 1.20 and greater than 0.8 and Zstd values were greater than -2.0 and less than 2.0 as established in the evaluation criteria. This indicates that there were no unexpected responses for items near ability levels for any student and that the INFIT statistics exhibit good fit and support the unidimensionality of the model. Based on these reliability indices and INFIT statistics, local independence was also supported. Evaluation of OUTFIT statistics show the MNSQ to be 1.65 which exceeds the critical value of 1.2. This information suggests a closer examination of individual item OUTFIT statistics is necessary.

The KR20 for the responses to the 25-item assessment was 0.39 which is less than the critical value of 0.80. The KR20-if-item-deleted and the point-biserial correlation for each of the items that were outside established criteria are presented in Table II together with Rasch item statistics to facilitate comparisons. The contribution of items: 1,5,8,10,11,15 were of particular interest because the point-biserial correlations are negative or the KR20 value increases when the items are deleted.

The Rasch item statistics (INFIT and OUTFIT) are presented in Table II in order of most misfit to least misfit. Item 1 is a suspect item because the OUTFIT MNSQ is 1.65 which is greater than the recommended 1.20, however the Zstd is 1.1 meeting the Zstd criterion. Likewise item 5 has an OUTFIT MNSQ of 1.31 and a Zstd of 0.2. The content of items 1 and 5 are seen in the Appendix. Since the Zstd is the standardization of fit values to a distribution with a mean of 0 and a variance of 1, both of these items deviate from the model only when values are not standardized. This suggests that the probability of these items affecting the interpretation of the instrument performance is unlikely. Further evaluation in relation to their overall contribution was performed to determine whether to retain the items or delete them from the assessment. Evaluation of item difficulty was obtained from the item logit value. Item 1 and item 5 are both difficult items as demonstrated by their high logit values 2.54 and 4.22 respectively. Unexpected responses demonstrated by the OUTFIT statistics suggest that some of the students who correctly answered these items were guessing or had special knowledge.

Evaluation of additional output provided by Rasch analysis provides information relating to individual item responses by student. The student responses to Item 1 show that it was correctly answered by five students. The item difficulty (logit value 2.54) was higher than the student’s ability for all five students as determined by the answer key. The highest level of ability of the five students was a logit measure of 0.5. The intended correct answer to item 1 was choice d) Pre-experimental. In examining the student’s responses to the other choices, six percent chose answer a) Database, 21 percent chose answer b) Crossover, 38 percent chose answer c) Experimental, and 29 percent chose answer e) Quasi-experimental. There is an observed variability of answers by the students for this item supporting that students may be guessing which response is correct.

Evaluation of student responses to item 5 shows that it was correctly answered by only one student. The item (logit measure 4.22) was correctly answered by one student of low ability (logit measure -0.41). The correct answer to item 5 was choice d) Compliance with manufacturer recommendations has been fully met. Choice a), The system demonstrates a low medication error rate, was selected by 85 percent of the students.

In addition to these two items not functioning as expected (as demonstrated by poor fit to the model), and that neither item 1 nor 5 as designed contribute much to the evaluation of student ability as evidenced by the absence of students who perform at the level of difficulty represented by those two items, 2.54 and 4.22 logits respectively. Thus items 1 and 5 were deleted from the assessment and the Rasch program was rerun.

The recalculated summary statistics for the Rasch analysis after deleting items 1 and 5 were reviewed. The item separation was 3.96 and item reliability was 0.94 indicating that items still create a scale and continue to support local independence. INFIT statistics continued to be within the evaluation criteria and support unidimensionality. Evaluation of the OUTFIT statistics showed the MNSQ max to be 1.22 which slightly exceeds the critical value of 1.2 and the Zstd max to be 2.3 which exceeds the critical value of 2.0. This information again indicates unexpected responses were detected and a closer examination individual item OUTFITS. The OUTFIT statistics in conjunction with INFIT statistics and reliability indices do however establish that the data fit the model and that information provided by the Rasch analysis is reliable and valid.

The KR20 for the responses to the 23-item assessment was 0.41. The KR20 increases slightly (0.01-0.02) for items 15, 10, 11, 8, or 6 if deleted from the assessment. The KR20-if-item-deleted and the point-biserial correlation for each of the 23 items are presented in Table III together with Rasch misfit statistics in order of most misfit to least misfit for each of the 23 items to facilitate comparisons.

Item 15 was a suspect item because the OUTFIT MNSQ is 1.22 however the Zstd is 0.5 meeting the Zstd criterion. The MNSQ is slightly greater than the recommended 1.20.

Table II -Rasch item misfits, point-biserial and KR20-if-item-deleted (initial analysis)

<table>
<thead>
<tr>
<th>Item</th>
<th>Logit</th>
<th>INFIT</th>
<th>OUTFIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>i1</td>
<td>2.54</td>
<td>1.05</td>
<td>1.65</td>
</tr>
<tr>
<td>i5</td>
<td>4.22</td>
<td>1.02</td>
<td>1.31</td>
</tr>
<tr>
<td>i15</td>
<td>2.17</td>
<td>1.05</td>
<td>1.20</td>
</tr>
<tr>
<td>i10</td>
<td>0.00</td>
<td>1.10</td>
<td>1.17</td>
</tr>
<tr>
<td>i11</td>
<td>0.17</td>
<td>1.08</td>
<td>1.14</td>
</tr>
<tr>
<td>i8</td>
<td>0.34</td>
<td>1.08</td>
<td>1.09</td>
</tr>
</tbody>
</table>

Values outside evaluation criteria or of interest are depicted in bold.
Likewise item 10 has an OUTFIT MNSQ of 1.21 and an OUTFIT Zstd of 2.3. Item 15 was identified as a difficult item (logit measure 2.46) and item 10 as a relatively easy item (logit measure 0.29).

Item 15 was correctly answered by seven students. The item difficulty (logit measure 2.46) was higher than the student’s ability for all seven students with only one student (logit measure -1.3) of the seven demonstrating a low logit ability measure of less than 1.0. It is this extreme value that caused the misfit statistic for the item. Examination of the item content and evaluation of the evidence from the student responses suggests that the item is very difficult which was consistent with the difficulty level that the item was developed to represent.

Evaluation of the student responses to Item 10 shows that it was correctly answered by 34 students. The item difficulty (logit measure 0.29) was approximately equal to or below the student’s ability for all three students. The item was flagged a misfit because of the unexpected responses of those three individuals. The item was designed to be relatively easy and the data support this with the logit measure of 0.29. Guessing is a possible reason for those correctly answering the item that demonstrated a slightly lower ability level. However, overall the misfit statistics are borderline supporting the idea that the item need not be excised.

The distribution map for the 23 items that represent the quality controlled assessment instrument is presented in Figure 1. Items are located on the right side of the scale. Student ability for each of the 80 students, is represented by a 2 for second-year students and a 3 for third-year students. It appears that the grouped pattern of second-year students have higher abilities than third-year students. Upon visual inspection there appears to be good item difficulty distribution with items covering areas where there is a measure of each student’s ability. The calculated item difficulty range is -1.90 logits to 2.46 logits which supports good item difficulty distribution based upon the range of the student’s ability as depicted in the item/ability distribution map.

However, there were gaps in the item difficulty hierarchy of the assessment. Construction of new items at item difficulty levels that might benefit the application of the assessment are depicted with arrows where the gaps in the items exceeded 0.30 logits (see Figure 1). The decision to add items in these areas is at the discretion of the instrument developer based on the intended application of the assessment.

Item 15 provides information that allows the evaluation of higher ability students. Item 10 provides assessment information in an area where there is no other item that measures that specific level of difficulty. Had there been other items on the map in the same location, an argument could be made to discard item 10. However, since both items 15 and 10 provide valuable information and the data provided by response patterns appears reasonable, valid arguments supported the retention of both items.

The students found the authentic assessment useful with 69 of 79 students (86.2 percent) reporting it helped them assess their learning needs. Seventy of the 79 students (87.5 percent) reported that this type of assessment would be useful in other areas of the curriculum. One item asked the students if they would recommend this type of assessment be performed before beginning rotations during the final year of the professional program. Sixty-eight of the 80 (85 percent) students agreed that this type of assessment should be performed prior to beginning rotations.

### GENERAL DISCUSSION

Traditional approaches of measurement based on CTT approaches do not support a measurement environment with the properties that permit the rigorous item evaluation necessary for the valid diagnosis of student, educator, and curricular deficiencies(43). The use of the KR20 and point-biserial correlation

---

<table>
<thead>
<tr>
<th>Item</th>
<th>Logit measure</th>
<th>INFIT MNSQ</th>
<th>INFIT Zstd</th>
<th>OUTFIT MNSQ</th>
<th>OUTFIT Zstd</th>
<th>Point Biserial</th>
<th>KR20 if item deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>i5</td>
<td>2.46</td>
<td>0.105</td>
<td>0.2</td>
<td>1.22</td>
<td>0.5</td>
<td>-0.05</td>
<td>0.42</td>
</tr>
<tr>
<td>iO</td>
<td>0.29</td>
<td>1.12</td>
<td>1.7</td>
<td>1.21</td>
<td>2.3</td>
<td>-0.09</td>
<td>0.44</td>
</tr>
<tr>
<td>ii1</td>
<td>0.46</td>
<td>1.07</td>
<td>0.9</td>
<td>1.13</td>
<td>1.3</td>
<td>-0.02</td>
<td>0.43</td>
</tr>
<tr>
<td>18</td>
<td>0.64</td>
<td>1.09</td>
<td>1.0</td>
<td>1.11</td>
<td>0.9</td>
<td>-0.04</td>
<td>0.43</td>
</tr>
<tr>
<td>i6</td>
<td>-1.09</td>
<td>1.08</td>
<td>0.6</td>
<td>1.04</td>
<td>0.3</td>
<td>0.01</td>
<td>0.42</td>
</tr>
<tr>
<td>i20</td>
<td>1.23</td>
<td>1.04</td>
<td>0.3</td>
<td>1.00</td>
<td>0.0</td>
<td>0.04</td>
<td>0.41</td>
</tr>
<tr>
<td>i3</td>
<td>-0.77</td>
<td>1.04</td>
<td>0.4</td>
<td>1.04</td>
<td>0.3</td>
<td>0.05</td>
<td>0.41</td>
</tr>
<tr>
<td>19</td>
<td>-0.31</td>
<td>1.02</td>
<td>0.3</td>
<td>1.03</td>
<td>0.3</td>
<td>0.09</td>
<td>0.40</td>
</tr>
<tr>
<td>i2</td>
<td>-0.83</td>
<td>0.98</td>
<td>0.2</td>
<td>1.03</td>
<td>0.2</td>
<td>0.12</td>
<td>0.39</td>
</tr>
<tr>
<td>i12</td>
<td>1.52</td>
<td>1.02</td>
<td>0.2</td>
<td>1.01</td>
<td>0.1</td>
<td>0.08</td>
<td>0.40</td>
</tr>
<tr>
<td>i6</td>
<td>1.38</td>
<td>0.99</td>
<td>-0.1</td>
<td>1.01</td>
<td>0.1</td>
<td>0.10</td>
<td>0.40</td>
</tr>
<tr>
<td>i2</td>
<td>0.13</td>
<td>1.01</td>
<td>0.1</td>
<td>1.00</td>
<td>0.0</td>
<td>0.10</td>
<td>0.40</td>
</tr>
<tr>
<td>i14</td>
<td>-1.45</td>
<td>1.00</td>
<td>0.0</td>
<td>0.98</td>
<td>-0.1</td>
<td>0.11</td>
<td>0.40</td>
</tr>
<tr>
<td>i8</td>
<td>-0.31</td>
<td>1.00</td>
<td>-0.1</td>
<td>0.97</td>
<td>-0.3</td>
<td>0.13</td>
<td>0.39</td>
</tr>
<tr>
<td>i3</td>
<td>V -1.80 \1.46</td>
<td>0.98</td>
<td>-0.1</td>
<td>0.89</td>
<td>-0.4</td>
<td>0.15</td>
<td>0.39</td>
</tr>
<tr>
<td>17</td>
<td>0.98</td>
<td>-0.1</td>
<td>0.94</td>
<td>0.94</td>
<td>-0.3</td>
<td>0.13</td>
<td>0.39</td>
</tr>
<tr>
<td>119</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.96</td>
<td>0.96</td>
<td>-0.4</td>
<td>0.18</td>
<td>0.38</td>
</tr>
<tr>
<td>125</td>
<td>0.24</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.95</td>
<td>-0.6</td>
<td>0.19</td>
<td>0.38</td>
</tr>
<tr>
<td>117</td>
<td>0.24</td>
<td>0.93</td>
<td>-1.0</td>
<td>0.72</td>
<td>-1.0</td>
<td>0.23</td>
<td>0.37</td>
</tr>
<tr>
<td>14</td>
<td>0.64</td>
<td>0.92</td>
<td>-0.9</td>
<td>0.87</td>
<td>-1.2</td>
<td>0.25</td>
<td>0.36</td>
</tr>
<tr>
<td>123</td>
<td>-1.37</td>
<td>0.92</td>
<td>-0.6</td>
<td>0.83</td>
<td>-0.9</td>
<td>0.25</td>
<td>0.37</td>
</tr>
<tr>
<td>i24</td>
<td>0.52</td>
<td>0.91</td>
<td>-1.1</td>
<td>0.92</td>
<td>-0.8</td>
<td>0.25</td>
<td>0.36</td>
</tr>
<tr>
<td>122</td>
<td>-1.90</td>
<td>0.91</td>
<td>-0.4</td>
<td>0.82</td>
<td>-0.7</td>
<td>0.24</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Values outside evaluation criteria or of interest are depicted in bold.
While the KR20 and Cronbach’s alpha are probably the most widely and easily calculated estimates of reliability, they lack precision and are usually not useful for evaluating small groups. An educator who has performed a Cronbach’s alpha or KR20 on a typical classroom assessment knows that correlation coefficients typically are less than 0.50. To improve reliability, additional items of similar structure and function need to be constructed. For example, using the Spearman Brown formula, we anticipated that we would need an assessment instrument 132 items in length to obtain a reliability coefficient of 0.80 using the reliability coefficient of 0.41 obtained from the final 23-item assessment. There are several problems with this. First, given the challenge of constructing good items, not many (if any) educators have the time to construct an additional 109 items. Secondly, if the items are constructed carelessly, the reliability could actually decrease. Furthermore, the “add more items to fix the problem” line of reasoning is a misrepresentation of the intent of Cronbach’s alpha and KR20. Unfortunately, it is easy to overvalue the meaning of reliability using coefficient alpha or KR20 simply by adding assessment items that are essentially images of one another. Assessments of greater length tend to have greater reliability, because the true score variance increases faster than error variance. AKR20 of 0.41 simply does not support lack of reliability in all instances.

This study provides a brief and nontechnical introduction to Rasch modeling. Rasch modeling provides an item function based analysis that serves as quality control of the construct that is not available with CTT; thereby, giving the assessment designer quality options to closely look at how each item is contributing to the overall assessment.

One of the advantages of Rasch modeling over CTT is realized by evaluating individual item misfits as a quality control mechanism for expected item function. With item 1, the intended correct answer, Pre-experimental represents a study design and/or terminology that students may be less familiar with and thus supports that the item is difficult. However, while initially it was thought that the study was a pre-experimental design, after further examination of the article, it was difficult to identify the design based on the information given. The study design was probably a before-after design using independent groups. The difficulty students had in selecting the study design type was also supported by the item OUTFIT MNSQ where guessing was a possible source of misfit. Given the difficulty in determining the design type, guessing by the students seems probable. This item (identify the study design) represented a component of the competency that should be considered critical in the evaluation of the literature. Reworking the item and responses would be necessary if the item is to function properly in the assessment.

Item 5 was not intended to be a difficult item. A possible explanation would be that students may have confused the word efficacy for effectiveness as a result of inattention or instrument burden. Regardless, the item OUTFIT suggests reworking the item so that it functions better if the assessment instrument is to be used again. In this case, a subjective examination without the benefit of statistics would suggest that the item does not work well.

Another advantage of Rasch modeling over CTT is realized if comparing point-biserial correlation coefficients to item misfits. For example, both methods detected items 1 and 5 as questionable. However, the Rasch model provided information favoring the retention of items 10 and 15 where the point-biserial did not. Additionally, the point-biserial correlation coefficients indicated that items 8 and 11 were questionable due to their negative values. This was primarily due to item difficulty for those persons of low ability. These two items performed well in the Rasch model. The use of Rasch modeling thus provided the assessment designer with a battery of detailed information with which to evaluate each individual item contribution to the assessment instrument.

Secondly, the Rasch model provided information suggesting that adding as few as three well constructed items to the authentic assessment would enable the instrument to better evaluate student performance. The item distribution map clearly depicts where additional items could possibly be of benefit. Benefit however is directly related to the use of the assessment. For example, if cutoff scores were used to make a judgment as to whether a pharmacy student would be allowed to begin rotations, the area surrounding that cutoff ability level should be adequate in the number and placement (in relation to difficulty level) of items.

Student “buy in” was demonstrated in this study. Of particular interest was the overwhelming support (85 percent of the students assessed) that this type of assessment be performed before the student begins rotations during the final year on the program. One student commented after the assessment, “Students should be required to show proof of competency before beginning rotations” another commented on the usefulness in providing the authentic assessment “I did not realize how much I did not know that I should.” There always exists the potential for desirability bias in this type of study. However, a concern of the researchers was that students were going to see this assessment as “another test” and thus a burden. Quite the opposite was found. Students showed support for the authentic assessment as demonstrated by their responses. Additionally, the distribution of scores and FIT statistics support that the students responded to the authentic assessment sincerely. Given the definition that authentic assessment is measuring knowledge and skills in a context that approaches that of the real world, the student performance level was not surprising since students do not have the opportunity to apply the content until their experiential rotations.

LIMITATIONS
It was assumed that respondents provided sincere responses to the assessment items. Rewards and punishments were not administered based on student performances on the assessment. Therefore, students had to rely on self-motivation to accurately respond to the assessment items. Measurement models are designed to provide only a representation of reality, not an exact one. The data collected represent only one point of assessment of an entire curriculum. The generalizability of this study is limited to the participants in the study program. However, the methods used in developing an authentic assessment and validating student performance can be applied to additional competencies and at other institutions.

RECOMMENDATIONS FOR FUTURE APPLICATION
The Rasch model is relatively easy to use and straightforward
to interpret when compared to other logistical models. Rasch analysis shows the student where he or she lies on the scale of competency. While this was not done with this study, in practice it would provide the student with precise information regarding what skills must be improved. The Rasch model accurately predicts the behavior of people and assessment items (43). This feature is especially useful to the student who desires to take responsibility for his or her learning.

Defining standards for minimal competency requires substantial time and commitment. Unfortunately, the scaling of students often relates more to subjectivity relating to measurement values than to real differences in performance determined by item difficulty. The subsequent application of competency standards to assessment would be relatively simple with Rasch analysis. In an interesting application of Rasch analysis, a program in a Chicago public school system used a cut score of 0.49 logits as the minimal level of ability to be certified for graduation and that student grades are reported in RITS (Rasch Logits) instead of letter grades like A through F (49).

One of the most promising applications of Rasch modeling over CTT is that it facilitates the creation and maintenance of an item bank (19). To obtain adequate item difficulty coverage when constructing new assessments, it would be a relatively simple task to pick and choose items from an established pool of known difficulty levels. Thus the classroom educator would be relieved of the burden of creating new assessments from scratch that are valid in their application. Once the item difficulty has been obtained, a pool of items may be collected for each particular level of difficulty or student ability. This would also allow a particular student to be assessed by using items obtained from an item bank that are within his or her ability. This prevents the student from devoting large amounts of time to items that are too easy or too difficult. When items are developed based on classical test theory, items cannot be selected from a pool of items that will meet specific measurement criteria (50). Taken one step further, items designed to measure competencies related to a specific domain could be selected from an item bank and used at several educational institutions that have similar educational learning outcomes.

To master material, the pharmacy student must be allowed to repeat the assessment after additional study. While this research used a cross sectional design to assess the usefulness of the instrument designed, it is hoped that the instrument would serve to facilitate learning and be incorporated for use in a longitudinal fashion. This would allow the student to continue to learn and retest until the skill or knowledge is mastered.

Authentic assessments should be closely related to curriculum frameworks. The establishment of institution specific learning outcomes provides an opportunity for action based curricular assessments. Unfortunately, curricular intentions do not provide assurance that students are integrating their learning over time and across multiple courses. To evaluate what students have learned and integrated, this type of authentic assessment could function as a milestone exam for curricular assessment. This study addressed only a small portion of the expected learning outcomes. It has been suggested, because of limitations of time and human resources, that evaluations should focus on dimensions in which students are deficient as determined by objective evidence (51). The further development of additional authentic assessments to evaluate other professional competencies is recommended.

Although it was beyond the scope of this study to assess curricular needs, it does provide useful information regarding student proficiency for the areas relating to critical analysis skills as related to the evaluation of professional/scientific literature. The item ability distribution map clearly showed that second-year students as a group outperformed the third-year students. A simple explanation would entertain that the second-year class was a better class. Another explanation would be that the third-year students were exhibiting knowledge decay. Additionally, it could be that the difference seen between the two groups was due to differences in how and what the students learned. This type of assessment coupled with Rasch analysis might also be of value when evaluating instructional experimentation. The data may be useful in providing specifics about what teaching styles are more conducive to student learning.

Rasch modeling provides information that can be used to evaluate performance from a curricular perspective. A closer look at those specific items on which the second-year students outperformed third-year students might shed light on how to improve the continuity of learning throughout the curriculum. This is supported in part by one third-year student's written comment “I do not feel that the information that I have previously learned concerning methodology can be applied without further secondary instruction.” The authentic assessment may serve as a baseline to make evidence based curricular-modifications as well as encourage responsibility for one’s own learning.

Further research including assessing student competency at multiple schools of pharmacy would be an application of this assessment that would offer a more global view of how students of pharmacy are performing in relation to their preparation for entering professional practice.

CONCLUSIONS

As stated by Williams and Redican “effective integration of standards, assessments, curricular, and instruction requires careful thought and planning” (52). There are numerous guides in pharmacy and education for creating competency statements and effective assessments that measure learning outcomes. Authentic assessments provide an approach that is close to the real world setting. It is essential that the approaches to evaluation and promotion of continuous assessment leading to student growth be valid and provide meaning. The perceived likelihood of students achieving an acceptable level of performance through personal efforts is directly related to this real world relevance (53). Thus the most valuable assessment techniques are those that provide benefits to both the student and the institution (11, 54).

This authentic assessment serves as a tool and example to help assist educators and students to accurately assess performance. Furthermore, accurate and valid authentic assessment of student performance was accomplished demonstrating evaluation of achievement with regard to desired learning outcomes. The validation procedure used a triangulation of all available information relating to the use and interpretation of the assessment instrument. Cronbach and Meehl supported this view in 1955 and it remains the focal point of assessment today (55).

The methodology used in developing this authentic assessment provides a template that should be used to assist in developing additional authentic assessments of other professional competencies. The use of authentic assessment in this study provided a mechanism in which the student had the opportunity to practice skills relating to competencies in a real-world
References

APPENDIX CONTENT OF ITEMS 1 AND 5

1. Which of the following study designs best describes this study?
   a. Database
   a. Crossover
   a. Experimental
   a. Pre-experimental
   a. Quasi-experimental

5. Efficacy would be established for the Medstation Rx system when
   a. The system demonstrates a low medication error rate
   a. The system meets with criteria developed by the pharmacy
   a. The system is accessible to all people who could benefit from it
   a. Compliance with manufacturer recommendations has been fully met
   a. Acceptance for use by pharmacy and therapeutics committee has occurred

References: