American Journal of Pharmaceutical Education **2004; 68 (2) Article 35.**

RESEARCH ARTICLES

Objective Curricular Evaluation: Applying the Rasch Model to a Cumulative Examination

JoLaine Reierson Draugalis, PhD,[a] and Terrence R. Jackson, PharmD, MSc[b]

[a]College of Pharmacy, University of Arizona
[b]College of Pharmacy, University of Illinois at Chicago

Submitted August 4, 2003; accepted September 28, 2003.

Objectives. The purpose of this study was to evaluate student and item performance and assess curricular strengths and weaknesses using a 65-item multiple-choice examination. Examination content included therapeutics, pharmacology, drug information and research design, nonprescription medications, and pharmacokinetics.

Methods. The cumulative examination was administered to 54 student volunteers prior to and following the final year of experiential rotations. Rasch analysis in combination with the Wolfe and Chui procedure was performed to evaluate content areas and compare pretest to posttest item and person values.

Results. These data supported Rasch model fit. Student logit values, converted to a 150-point scale, showed mean measures of 94.0 and 97.8 for the pretest and posttest, respectively, demonstrating significant improvement ($t = 3.20$, $p < 0.01$). Calibration values showed that 12 items became easier and 13 items became more difficult.

Conclusions. The data provide evidence of student learning during the fourth experiential year of the doctor of pharmacy program. Additionally, this analysis provided information that may be used to guide curriculum changes.

Keywords: curriculum, student evaluation, Rasch model, examination

INTRODUCTION

Objective tracking of student performance during the fourth experiential year and assessing curricular strengths and weaknesses in a doctor of pharmacy program is challenging to perform and seldom reported in the literature. Typically, pharmacy programs use subjective data provided by rotation preceptors to evaluate student performance during this experiential year. Even then, the collation of these data are segmented or inconsistent and offers little in the way of valid evidence for objectively assessing student performance or the curriculum. When feedback about the curriculum as a whole does occur, it is most often provided anecdotally from student exit questionnaires. This summative information may be useful, but it is imprecise and insufficient for making recommendations to the curriculum committee regarding changes that lead to improvements in expected learning outcomes.

Corresponding Author: JoLaine Reierson Draugalis, PhD, Professor and Assistant Dean, College of Pharmacy, The University of Arizona P.O. Box 210207, 1703 E. Mabel, Tucson, AZ 85721. Tel: 520-626-5301. Fax: 520-626-4063. Email: draugalis@pharmacy.arizona.edu.

Cumulative written examinations, eg, multiple choice questions (MCQ), are the most appropriate and efficient mechanism for assessing student cognitive knowledge and are also capable of testing higher-order cognitive levels (ie, decision-making, problem-solving).[1,2] In medical school curricula, content-rich cumulative-knowledge examinations are administered following completion of the didactic component but prior to entering a clerkship as a required step toward licensure.[3] In pharmacy, this cumulative knowledge examination occurs as a requirement for licensure only after completion of the degree program, when the opportunity for formative assessment has passed. In a study reported by Ryan and Nykamp in 2000, only 6 schools/colleges of pharmacy in the United States reported using MCQ format for cumulative cognitive knowledge assessment. Of those, 1 school used the examination to provide feedback to the students and 1 school used data from a cumulative examination as a tool for evaluating the curriculum.[4] Objective assessment of student knowledge and synthesis of student achievement maps or student-curricular maps have been useful in providing a mechanism for interpreting student achievement

or assessing curricular objectives outside the health-professions arena.[5,6] However, mapping student achievement through cumulative examinations to evaluate content-specific areas provided throughout the pharmacy curriculum has not been reported. Relevant to all assessment but most important in testing is validity.[7] That is, the accurate and meaningful interpretation of test results and the inferences based on them must be a primary consideration.[8,9] Outside national board examinations, the gathering of evidence to support construct validity regarding the performance of cumulative examinations via item functioning is rare.[10,11] The 2 major threats to validity, construct irrelevant variance (eg, poorly crafted items, guessing, item bias) and construct underrepresentation (too few items, trivial questions), are especially important considerations in MCQ testing.[8,12] Fortunately, the richness of psychometric characteristics inherent with good MCQ items and objective measurement techniques makes the evaluation of these threats practical.[13,14]

The purpose of this study was to evaluate student proficiency and assess item performance of a cumulative MCQ-format performance assessment in a doctor of pharmacy program. Specifically, this study utilized Rasch analysis in a pretest/posttest design to evaluate student cognitive knowledge for the content areas of therapeutics, pharmacology, drug information and research design, over-the-counter medications, and pharmacokinetics. Thus, data were used to assess the validity of the inferences provided by the performance of the assessment instrument and to objectively evaluate curricular strengths and weaknesses via the content of individual items.

METHODS

Study Design

This study used a single group pretest-posttest design. Initially, a 75-item multiple-choice-question performance assessment was assembled by the Assistant Dean for assessment and evaluation, with a variety of items suggested by course coordinators. This project used standards and guidelines from the accreditation standards of the American Council on Pharmaceutical Education (ACPE) and expected learning outcomes from the American Association of Colleges of Pharmacy (AACP) as guiding principles for development of specific institutional competencies delineated in the College's "Outcomes Expected of a Graduate" document.[15-17] These documents served to guide the curriculum as a valid source for selecting content areas with which the MCQ format could be used effectively. In addition, the number of items covering each content area was determined based on the amount of didactic time students spent on the subject during the first 3 years of the professional program while keeping in mind the effect that instrument burden has on an examination that is completed on a voluntary basis.

One consideration of validity refers to the sampling adequacy of the content area being measured. Content validity may be supported through the use of *Marzano's Taxonomy*, which is a design based on information gathered since the introduction of *Bloom's Taxonomy*, and claims to resolve many of the difficulties associated with *Bloom's Taxonomy*.[18] *Marzano's Taxonomy* also demonstrates better alignment between the cognitive level intended to be tested in the original construction of the assessment item than does *Bloom's Taxonomy*.[19] This study used *Marzano's Taxonomy* as a cognitive template with which to build the assessment blueprint. Additionally, to make accurate interpretations regarding the outcomes of the assessment tool, it is imperative to evaluate how the tool was constructed and subsequently used. Haladyna and Downing's 31-item checklist was used in identifying items that could further benefit from finetuning.[20]

The 75-item instrument was administered as a pilot test to 54 third-year students 2 mo prior to entering their fourth-year experiential rotations.

The Measurement Model

The Rasch dichotomous model was selected to evaluate the data provided by the instrument because it provides objective evidence, when data fit the model, that all items measure the same construct (ie, unidimensionality) and produce additivity of measures (ie, true interval level data), and that the probability of the student correctly responding to an item does not depend on the other items in the assessment (ie, local independence).[21,22] Four Rasch criteria were used to provide additional evidence to support construct validity. These included mean square (MNSQ) INFIT and OUTFIT statistics, separation reliability, and the item distribution map.[14]

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. These statistics provide empirical evidence to detect when the following occur: (1) an item is not part of the same dimension being measured, (2) an item is not understood, and (3) a specific response demonstrates that the student

did not take the assessment seriously (ie, acquiescent response bias or guessing) or had special knowledge (ie, true special knowledge or the examination had been compromised).

Construct irrelevant variance was evaluated in this assessment using MNSQ INFIT and OUTFIT statistics. The Rasch model provided valuable information for detecting items in the original 75-item assessment instrument that performed poorly, contributed to construct irrelevant variance, and/or produced measurement redundancy.

Item difficulty is described on a measurement 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. 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. Because logits are reported in positive and negative values, logit measurement units were rescaled from 0–150 to enable the reader to interpret its meaning in positive units. The new scale is similar in structure to that used by the National Association of Boards of Pharmacy (NABP).

Item separation is the distance in logits between items of varying difficulty. Item reliability is the estimate of reproducibility of item placement within the hierarchy of difficulty across students of differing abilities. The item separation index is expressed in standard error units as calculated by dividing the adjusted item standard deviation by the average measurement error. The reliability of item separation is determined by the extent that item calibrations are sufficiently spread out to define distinct levels of ability measured in logits.

An item distribution map is constructed to show the distribution of the persons and items on the same measurement scale (see Figure 1). The ability of the Rasch model to demonstrate the relationship between a person's ability and item difficulty on the same measurement scale is a property inherent to the model when the data fit the model. The scale measuring the construct is laid out vertically with the most able persons and most difficult items at the top.

The Wolfe and Chiu procedure, modified for dichotomous responses, was used to compare pretest to posttest items and person values.[23] The Wolfe and Chiu procedure uses an anchoring strategy for measuring change in person ability measures and item calibration values over time. That is, it provides a method for determining if the observed differences were the result of change in person measures due to the intervention and not due to changes in the measurement situation, regression to the mean, maturation, or experimental dropout. Thus, a potential threat to the internal validity of the interpretations is eliminated.[23]

Rasch person logit values were converted to a 150-point scale similar to that used in NABP examinations to explore student performance. This was done to both facilitate interpretation of student ability in a manner with which the pharmacy community is familiar and to use values positive values.

Subjects

The subjects used in this study were students enrolled in the doctor of pharmacy degree program at the University of Arizona College of Pharmacy, Class of 2002. The Offices for the Protection of Research Subjects at the University of Arizona and the University of Illinois-Chicago granted approval for this research project. The instrument was administered to student volunteers prior to and following the final year of experiential rotations. Responses from the initial 75-item administration (given during regularly scheduled therapeutics class time) were retained for the 65 items of the instrument that met model requirements. Only the 65-item instrument was administered to students following the final year of experiential rotations during senior week activities. This was a secure, paper-and-pencil examination that was given via group administration. It was important to the authors that the examination be secure so as not to compromise the validity of the inferences obtained from responses to the assessment. The following factors were also considered for test administration: time issues, instrument burden, and standard conditions. Students were given sufficient time to complete the examination. The focus of this examination was not on speed, but rather on knowledge so as to avoid negatively effecting validity and reliability regarding the performance of the assessment.

Statistical Analysis

Data (containing dichotomous responses to items of the assessment instrument) were entered into a data file with MS-DOS and then input into Winsteps version 3.37 (Mesa Press, Chicago, Illinois) to calculate statistics for the Rasch model.[24] SPSS statistical analysis system version 11.0.1 for Windows (SPSS Inc, Chicago, Illinois) was used to calculate *t*-tests for evaluating pretest and posttest assessment data.

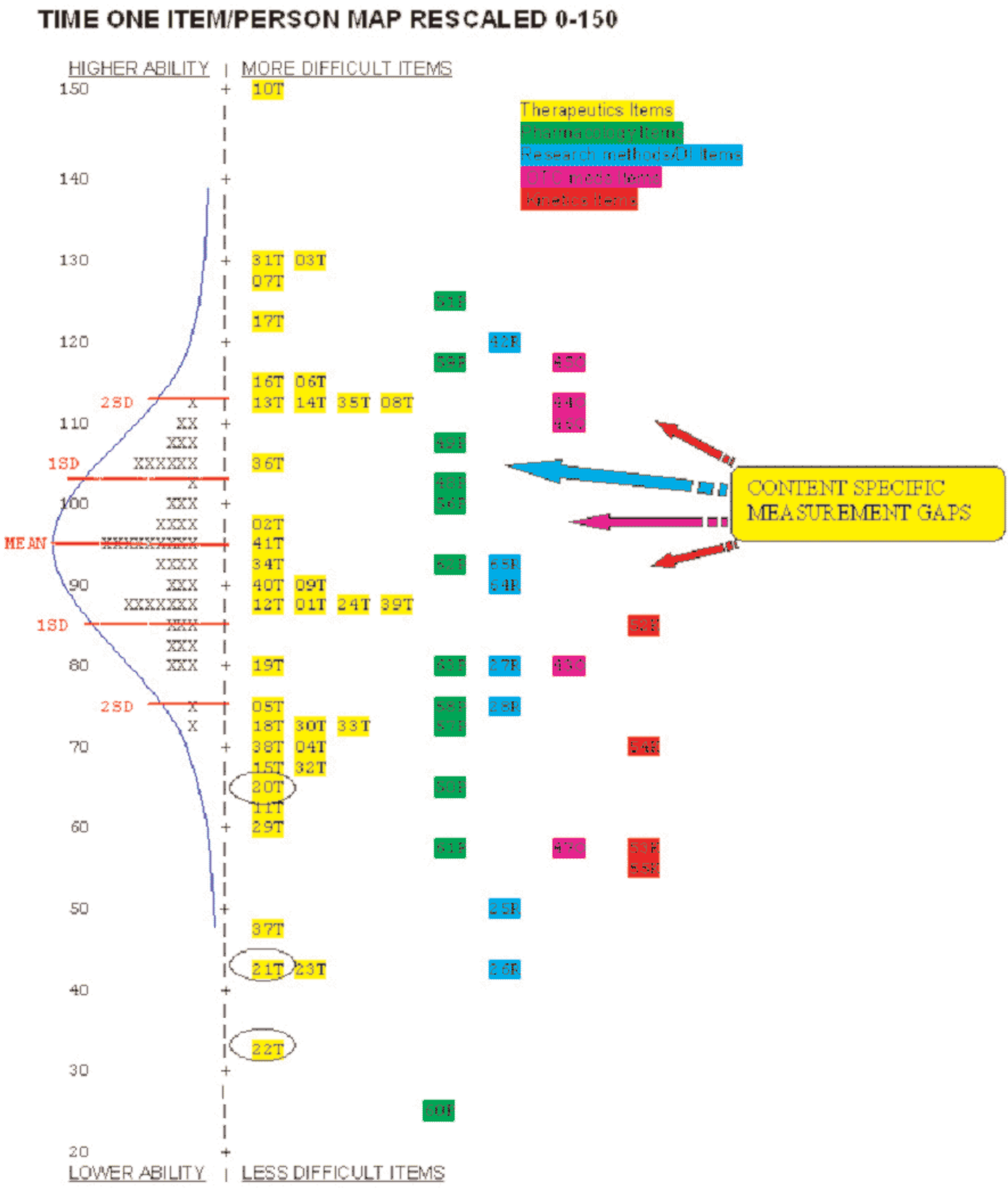

Figure 1. Time one item/person map for the 65-item assessment instrument showing item content distribution and student targeting on a scale from 0–150.

RESULTS

The assessment instrument was completed by 54 students (100%) in March 2001, prior to beginning their experiential rotations, and again in May 2002, following completion of their final year of experiential rotations. Rasch analysis was performed on the initial 75-item assessment, and subsequently10 items were deleted from the instrument because of item misfit statistics (ie, MNSQ Outfit greater than 1.2), measurement redundancy (ie, MNSQ Outfit less than 0.8), or low content balancing (ie, 1 remaining law item). The remaining 65 items were found to function unidimensionally to measure the underlying pharmacy proficiency construct. The final 65-item assessment blueprint consisted of 37 items related to therapeutics, 12 for pharmacology, 7 for drug information and research design, 5 for over-the-counter medications, and 4 for pharmacokinetics.

Model Fit, Unidimensionality, and Construct Representation

Evaluation of INFIT and OUTFIT statistics for the items in the assessment instrument showed that MNSQ values were less than 1.2 and greater than 0.8. Thus, these data exhibited good model fit and supported the unidimensionality and local independence requirements of the model.

The separation index (ie, the extent that items are sufficiently spread out to define distinct levels of ability) for the 65-item assessment instrument was 4.7. That translates to an item reliability of 0.96 (ie, the estimate of reproducibility of item placement within the hierarchy of difficulty across students of differing abilities), indicating that the items created a variable that was well spread out and that item placement along the scale was reliable. The separation index for the 54 students was 1.0, which represents a person reliability coefficient of 0.48 (analogous to Cronbach's alpha).

An item distribution map was constructed to display visually the distribution of the student abilities and items on the same measurement scale in units from 0 to 150 (Figure 1). The scale measuring the pharmacy proficiency construct was laid out vertically with the most able persons and most difficult items at the top. This map shows visually the relationship between student performance and item difficulty. The left-hand column is used to locate student ability and the right-hand column is used to locate item difficulty. The items were color coded to facilitate identification of the specific content area that each item represented. Overall, the cumulative assessment instrument demonstrated an item difficulty distribution that was well targeted to the student population. That is, from a measurement perspective, each student's ability, as depicted by each "X" in the normative distribution provided on the right side of Figure 1, was assessed well by items with calibration values in the same region of the measurement scale. Specifically, the content areas related to therapeutics and pharmacology were adequately represented and distributed throughout the measurement continuum. However, while the content areas related to research design, drug information, nonprescription medications, and pharmacokinetics were well distributed in the item difficulty hierarchy, there were small measurement gaps in the higher ability area of the assessment for these content areas.

Comparison of Student Ability Measures

The group means for student ability measures were 94.0 (±9.4) and 97.8 (±9.3) on a scale of 0–150 for the pretest and posttest, respectively. Dependent student's *t*-test demonstrated the difference in group means to be a significant improvement from pretest to posttest ($t = 3.20$ $p < 0.01$). Comparison of individual student measures (ie, pretest to posttest) showed that proficiency improved for 33 of 54 students and was maintained for 6 of 54 students. Fifteen students showed a decline in performance. One of those 15 students showed a statistically significant decline in ability measures (ability measure 81.1 to 55.1, $t = 2.35$, $P < 0.05$).

Comparison of Item Calibration Values

Analysis of data showed that item calibration values for 39 of 65 items were invariant from pretest to posttest (ie, item calibrations were stable). Thirteen item calibration values showed changes that were greater than 0.3 logits in difference from pretest to posttest, which indicated that these items became easier for students to answer correctly in the posttest (Figure 2). Of the 13 items, 10 related to therapeutics, 2 related to nonprescription medications, and 1 related to pharmacology. Furthermore, 7 items exhibited substantial calibration changes (ie, greater than 0.5 logit difference) from pretest to posttest. Five of these 7 items related to therapeutics, of which 1 item demonstrated a significant calibration value change (-0.46 to -1.82 logits, $t = 2.46$, $P < 0.05$). Two of the 7 items related to nonprescription medications.

Conversely, 13-item calibration values showed changes that were greater than 0.3 logits in difference from pretest to posttest, which indicated that these items became more difficult for students to answer correctly in the posttest (Figure 3). The content of 5 of these items related to therapeutics, 3 related to pharmacology, 2 to

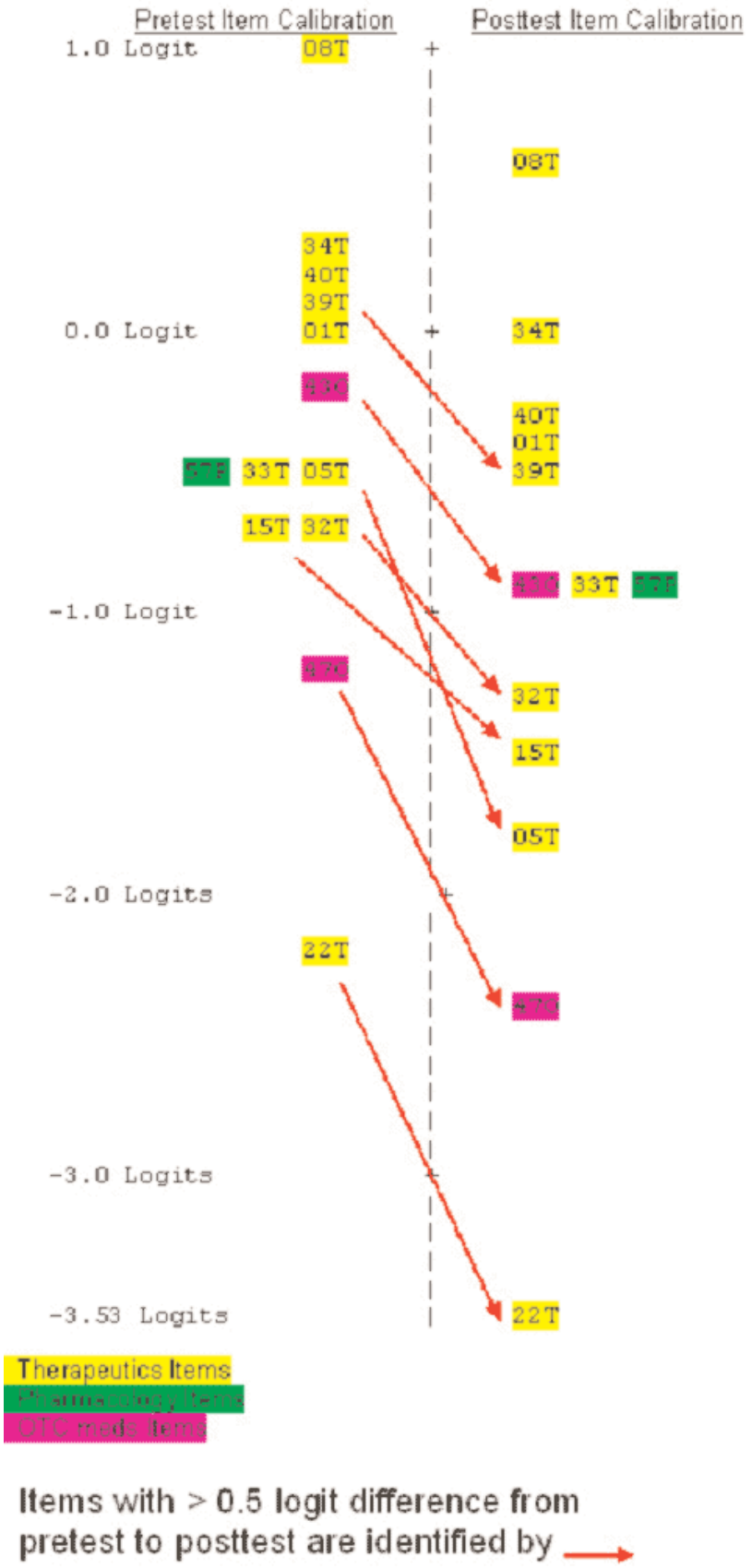

Figure 2. Comparison of item calibration values for items that became easier for students.

research design, and 3 to pharmacokinetics. Of the 13 items, 4 exhibited substantial calibration changes (ie, greater than 0.5 logit difference). Two of these items related to kinetics, 1 related to pharmacology, and 1 related to research design. The item relating to research design demonstrated a significant calibration value change (-0.46 to +0.88 logits, $t = 3.32$, $P < 0.01$).

DISCUSSION

To make accurate interpretations regarding the outcomes of any assessment tool, it is imperative to correctly evaluate how the tool was constructed and subsequently used. That is, validity issues must be considered part of the design template when developing an assessment instrument.[7,25] Initial evidence to support validity for the performance of the cumulative pharmacy assessment used in this study was initially provided when the following occurred: (1) items were selected and/or developed using the competencies delineated in the College's document, "Outcomes Expected of a Graduate," (2) *Marzano's Taxonomy* was used to verify item cognitive appraisal, and (3) Haladyna and Downing's item development checklist was used as quality control for item structure.[15,18,20]

Subsequently, Rasch analysis was used as quality control to identify items that contributed to providing construct irrelevant variance and 10 items were omitted from the original 75-item pilot instrument. The loss of these 10 items was primarily due to the finding that 4 items relating to law content most probably represented a construct different than that of the rest of the assessment instrument. The authors were not surprised by this finding. There were 4 items related to therapeutics and pharmacology, which collected information that was of little value from a measurement perspective. Essentially, the correct answer to these items could be selected without having extensive knowledge about pharmacy. While it could be argued that responses to the content of these items may still be useful, the authors decided that the reduction in instrument burden was the larger gain.

Sampling Adequacy

Another consideration of validity refers to the sampling adequacy of the content area being measured. This was accomplished in part by taking into consideration the depth of cognitive knowledge (ie, comprehension, analysis, knowledge utilization). For example, the assessment included the use of several small case studies requiring various levels of cognitive skill as defined by *Marzano's New Taxonomy*, to be able to answer correctly.[18] Items 20T, 21T, and 22T accompanied one such clinical case and represented the use of knowledge utilization, analysis, and comprehension, respectively. These items also varied in hierarchical difficulty as determined by Rasch analysis (see circled items in Figure 1).

Additionally, the evaluation of construct underrepresentation was performed by evaluating the measurement distance between item calibrations, which is depicted visually on the item and person distribution map (Figure 1). This map clearly shows that the content areas related to therapeutics and pharmacology were adequately represented and well distributed throughout the measurement continuum. However, while the content areas related to research design, drug information, nonprescription medications, and pharmacokinetics were also well distributed in the item difficulty hierarchy, 3 small measurement gaps for these content areas were identified (identified in Figure 1 using arrows). While the item sampling was originally guided in part by the amount of didactic time students spent on the subject during the first 3 years of the professional program, from a measurement perspective the development of additional educational items in these areas would improve the ability of the instrument to differentiate levels of performance for those with higher ability and better detect differences over time with better precision.

Population Targeting

Appropriate item targeting of student ability is an important consideration in efficiently measuring student ability and obtaining the precision necessary to detect change over time. For dichotomous data, the maximum information function of an item occurs when the probability of correctly answering the item for a given person, $p = 0.5$ as information = $p(1-p)$.[26,27] That is, the closer the person measure to the item calibration value (ie, well targeted), the more information that item contributes to the precision of the measurement of the person's ability. The greater the distance between the person measure and the item calibration value, the less efficient the item becomes. Thus, assessment efficiency is lost and a greater number of items are needed to obtain an ability measure of comparable precision. This cumulative pharmacy assessment demonstrated good targeting to the student population, which can be observed visually in Figure 1 by comparing the distribution of students on the left side of the map to the distribution of items on the right side of the map. Because these items were designed to be difficult and were well targeted to the student population, they maximized the amount of information provided by the assessment. This allowed for efficiency in the number of assessment items administered, which enabled the detection of

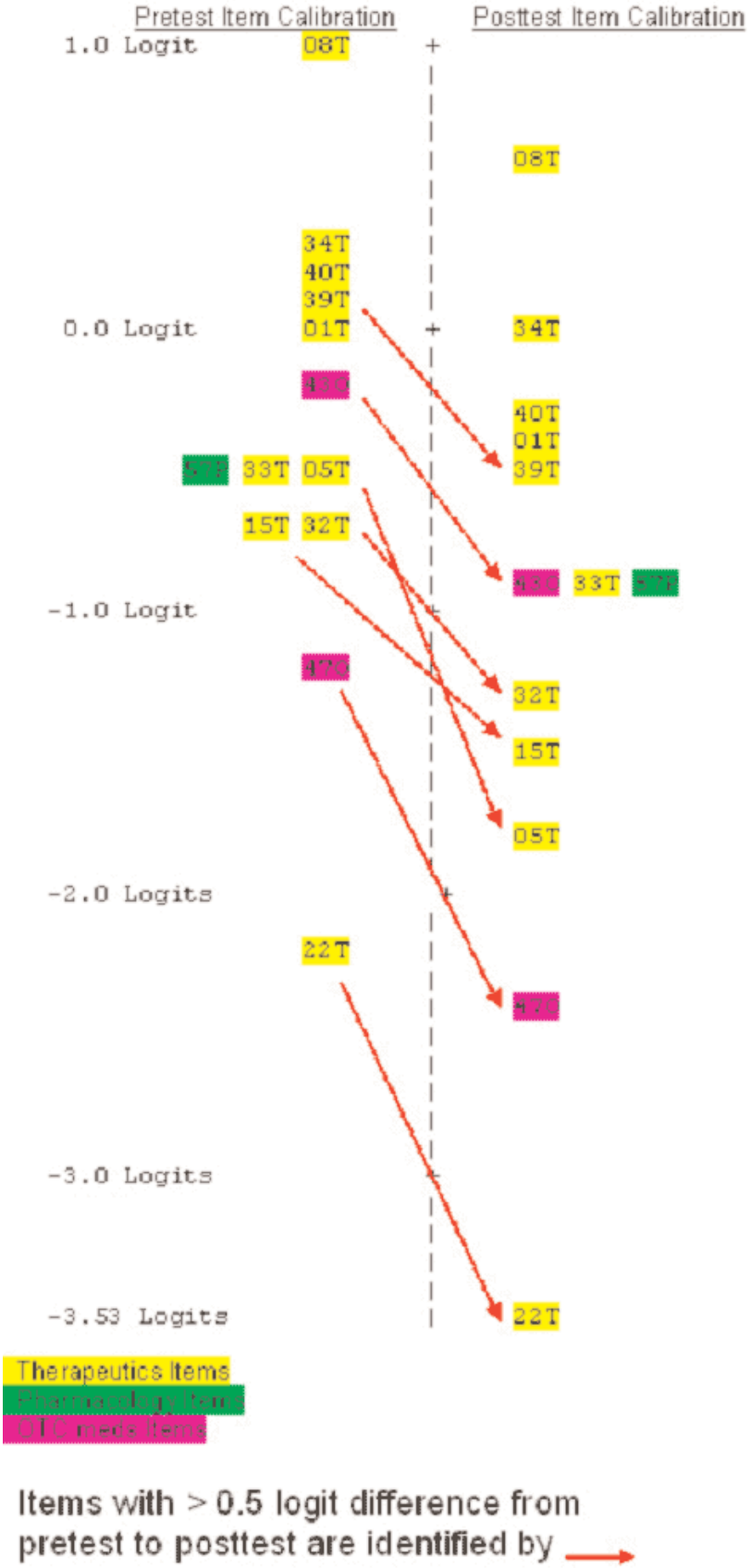

Figure 3. Comparison of item calibration values for items that became more difficult for students.

change over time in the variables related to knowledge of pharmacy. Similarly, it is this measurement precision and targeting that allows for differences to be detected even when as few as 30 subjects are used.[28]

Construct Validity

The use of Rasch modeling as a mechanism for quality control led to the deletion of 10 items not contributing to the good measurement properties of the instrument. Thus, allowing a reduction in instrument burden by 13% without loss of measurement precision. The final 65-item assessment blueprint exhibited unidimensional characteristics and model fit (ie, INFIT MNSQ and OUTFIT MNSQ greater than 0.8 and less than 1.2). Additionally, the item separation reliability (4.68, 0.96) provided additional evidence to support that the 65 items reliably provided information with which to assess the population for which it was intended.

Comparison of Student Ability Measures

The Rasch model combined with the Wolfe and Chiu procedure provided interval level data that enabled the use of parametric statistical procedures for comparison of responses from pretest to posttest. This method also allowed the authors to determine whether the observed differences in person ability measures were due to the intervention (ie, the experiential year) and not to changes in the measurement situation. In this study, data were transformed to a user-friendly scale from 0 to 150 to facilitate interpretation in a manner similar to that used by the NABP for licensure. Data show that the sample distribution is homogenous, with little variability among students. Coupled with the small sample size and relatively few items in the assessment instrument (ie, 65) the reliability coefficient for students was 0.48, which is typical under these conditions. However, because of the measurement precision the Rasch model provided, detection of differences among the students was possible.

The comparison of group means for student measures (ie, 94.0 to 97.8) demonstrated statistically significant improvement ($t = 3.20$ $p < 0.01$) in proficiency. Thus, supporting that learning among students had occurred during the fourth experiential year. This finding in and of itself provided a substantial contribution for empirically evaluating student learning outcomes that occurred over the fourth experiential year of the program.

However, Rasch modeling also provided information about students evaluated at the individual level. Individual student measures showed that proficiency improved (ie, 33 of 54 students) or was maintained (ie, 6 of 54 students) for the majority of students. The finding that 15 of 54 students did not show an improvement after completing their experiential year was bothersome. This may be evidence of a disconnect between the didactic and experiential component of the curriculum, or a possible disconnect between the content areas examined and competencies achieved during experiential training. This may also represent a limitation of MCQ testing. That is, additional methods for assessing performance of skills beyond the measurement of cognitive knowledge is an important process that many health professional programs are currently pursuing.[29,30]

However, the 5 highest scoring students on the pretest were also in the 25th percentile of their class. The consistency between these 2 measures provided additional validity that the content of the cumulative assessment was aligned with that of the curriculum in providing evidence of student learning. Only 1 of the 5 highest scoring students on the posttest was in the top 25th percentile of class ranking. This finding was not surprising because the type of learning students were exposed to during the fourth experiential year was rooted in adult learning theory and supported personal growth via self-directed learning, whereas those excelling in the pretest were reflecting their ability to learn in a didactic setting.

Evaluation of individual student ability measures showed there were 2 students who performed at proficiency measures of less than 75 in the posttest. The lowest performing student provided an ability measure that was significantly less on the posttest than the pretest ($t = 2.41$, $p < 0.05$). The second of these students also performed at a measure that was less for the posttest than the pretest, but not statistically significant. The interpretation for their performance was closely scrutinized because the examination was voluntary and no stakes were involved. On low-stakes examinations, high performance probably correlates with high knowledge, but low performance may not correlate with low knowledge. However, examination of person fit statistics, a great advantage of performing Rasch analysis, showed that the individual MNSQ INFIT/OUTFIT scores were 1.1/1.1 for one student and 1.0/1.0 for the other. That is, the pattern of responses to individual items was consistent with that of having lower ability and not characteristic of a person who would have responded haphazardly. These values supported that the 2 students took the assessment seriously and their ability measures accurately reflected their proficiency. This is of particular concern because this evidence suggested that had a mechanism for tracking learning over the fourth experiential year been introduced earlier, these students

may have been identified and an appropriate intervention developed that was targeted to student deficiencies. Additional evidence supporting the validity of the interpretations for these 2 students would have included an evaluation of these students' performances on experiential training. However, the nature of this study prohibited the disclosure of this information.

There were no raw score passing standards associated with this examination. This examination was administered from a measurement perspective. However, the use of a 0 to 150 scale (as used with the NABP examination on which a score of 75 is considered passing) allows the reader to visualize the validity in arriving at the demonstrated student ability measures. Given the item difficulty hierarchy, it would be relatively easy to assign a pass score to this assessment using Angoff's method.[31,32] Additionally, the examination was a low stakes examination. Thus, students did not prepare for it as they would a high stakes examination. This worked to the advantage of the researchers as the authors were interested in what students learned and retained as a result of their fourth year experiences.

Comparison of Item Calibration Values

Rasch analysis also provided information with which to evaluate individual item characteristics and contributions to the measurement properties of the instrument. Because of this, these data provided detailed information with which to evaluate content areas represented by each item and enabled the evaluation of content proficiency from a curricular perspective regarding performance during the fourth experiential year.

The item separation and reliability for the assessment instrument, 4.7 and 0.96, indicated that the items created a variable that was well spread out and that item placement along the scale was reliable. Typically, item calibration values remain relatively stable (ie, are invariant) when used to track changes in performance over time (eg, longitudinal studies). The Wolfe and Chiu procedure, used for invariance evaluation, was useful in that it allowed the researchers to determine the extent to which item calibrations were stable across the 2 measurement occasions. Differences in calibration values over time demonstrated by specific items yielded valuable information regarding content areas that were subject to normal knowledge decay, and more importantly, about the reinforcement of content areas during the fourth experiential year. Item calibration stability, demonstrated by differences less than 0.30 logits, can be expected for most variables.[28] There were 39 items demonstrating calibration value differences less than 0.3 logits and they were used as anchor items for the Wolfe and Chiu procedure. Differences in item calibration subsequent to using the Wolfe and Chiu procedure of greater than 0.3 logits and 0.5 logits were used to evaluate the effect of the fourth experiential year on curricular content areas. In particular, item calibration values reflecting differences that reached statistical significance were considered areas important for further inquiry.

There were 13 items that exhibited item calibration shifts of greater than 0.3 logits, indicating that these items became easier for students to correctly answer (see Figure 2). Seven of those 13 items demonstrated shifts in calibration values greater than 0.5 logits. The shifts in calibration values, especially those with calibration shifts of greater than 0.5 logits, provided evidence that those specific areas of the curriculum had been reinforced during the fourth experiential year. Closer evaluation of these content areas showed that overall the 13 items reflected content related to therapeutics (ie, 10 of 13 items). Two items represented content related to nonprescription medications and 1 related to pharmacology. The calibration shifts for these areas intuitively make sense and support the validity of the interpretations because the use of content related to therapeutics is consistent with areas students are continuously exposed to during the fourth experiential year. An overview of these areas reaffirmed that these content topics were indeed reinforced during the fourth experiential year. Of particular interest was the calibration value shift for item 5, which was significant ($t = 2.46$, $p < 0.05$). This item specifically related to an international normalized ratio (INR) for therapeutic anticoagulation in a patient following a myocardial infarction. Essentially, students had to be able to use different therapeutic ranges and determine which range was appropriate for which condition in order to respond to the question correctly. The magnitude of this calibration shift supported the development of higher-level cognitive function (eg, content mastery) for students in this area of therapeutics.

There were 13 items that exhibited negative item calibration shifts of greater than 0.3 logits, indicating the items became more difficult for students to correctly answer (see Figure 3). Only 4 of those 13 items demonstrated shifts in calibration values greater than 0.5 logits. Three of the 13 items specifically related to the use of pharmacokinetics. The calibration value shift of greater than 0.3 logits but less than 0.5 logits for the items related to pharmacokinetics was not surprising because students completed their second semester of this coursework at the end of the second semester of their third year in the program. Taking the pretest while enrolled in the clinical pharmacokinetics course allowed students to perform at

what was likely their peak in this content area. The ability of the Rasch model to identify the decrease in student ability for content areas related to pharmacokinetics again supported the validity of the model to detect change over time. However, differences greater than 0.5 logits for 2 of these items is also of concern because anecdotal reports suggest that pharmacokinetics is infrequently reinforced during the fourth experiential year.

A second major area of concern was the decreased performance related to information about research design (ie, items 28 and 65). Item 28 demonstrated a calibration value shift that was significant ($t = 3.32$, $p < 0.01$). While both of these items related to content that represented skills needed to critically evaluate published scientific literature consistent with the practice of evidence-based medicine, the magnitude of the calibration shift for item 28 warranted a recommendation to the curriculum committee to provide a curricular intervention aimed at improvement in this area.

The Rasch model only identified changes over time for student ability and item performance. The researcher must then evaluate the importance of the changes. One question this raises is whether there is a need to continue basic science exposure through the experiential year. The pharmacokinetics information was provided to the curriculum committee to further evaluate whether the final item calibration values for these items represented acceptable knowledge use and application by the students upon completion of the fourth experiential year, or if a recommendation for curricular change was warranted.

For the areas related to research design, initial evaluation for constructing a recommendation for improvement seemed relatively simple to implement. For example, the provision of more opportunities for students to practice these skills (eg, journal club) was a logical recommendation. However, further evaluation suggested that student experiences that would include the use of critical analysis skills extending beyond the scope of a journal club may be difficult because this is an area in which many preceptors may be less comfortable in providing guidance to students. Therefore, the empirical evidence provided by this study supported the recommendation for curricular improvements in the content area of research design that included an intervention aimed at continued professional development for preceptors.

CONCLUSIONS

The methodology used in the development of this cumulative pharmacy assessment provides a template that allows student learning to be assessed and curricular strengths and weaknesses evaluated. Specifically, this study used a cumulative MCQ cognitive knowledge assessment that was administered to volunteer pharmacy students prior to and following the final year of experiential rotations. While the administration of the cumulative assessment instrument occurred at only 1 school, the methodology is generalizable to other colleges and schools of pharmacy.

Evidence supporting the validity of the method used in this study was initially provided by content matching assessment item content to course material and the College's "Outcomes Expected of a Graduate" document.[15] Rasch analysis facilitated the provision of construct validity evidence, which was essential in establishing the level of precision that was used to evaluate the results of this study. However, as previously discussed, the development of additional items in the higher ability range for the content areas of research design, drug information, nonprescription medications, and pharmacokinetics would improve the measurement precision of this assessment. Rasch analysis also made available student ability measures and standard errors that provided evidence of student learning that occurred and/or was maintained during the fourth experiential year.

Data in this study also provided specific information that may be used to guide curriculum changes and provide guidance for experiential preceptors. The identification of curricular strengths (ie, content area related to therapeutics) was consistent with the expected learning outcomes of the college. The identification of content areas where curricular recommendations, consistent with continuous quality improvement, in the area of research design were deemed appropriate, were based on objective measurement and empirical data.

Rasch analysis provided a quality control function that established assessment item performance that is consistent with high stakes examinations used by licensure boards. While the researchers are aware of only a few colleges of pharmacy that have faculty members who are proficient in the area of objective measurement (ie, Rasch analysis), most schools/universities have faculty members in other departments (eg, educational psychology, physical therapy) who are proficient in this area. Additionally, a list of Rasch consultants may be found at the Web site for the Institute for Objective Measurement.[33]

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues at the University of Arizona College of Pharmacy for providing input on item selection and development. Particularly Dr.

Brian Erstad for overseeing the contributions of the therapeutics-related items. Lastly, we acknowledge the encouragement and valuable input of Dr. Everett V. Smith, Jr, Associate Professor at the University of Illinois at Chicago College of Education. His expertise in psychometrics and objective measurement was especially helpful in producing this manuscript.

REFERENCES

1. Downing SM. Assessment of knowledge with written test forms. In: Geoff RN, Van der Vleuten CPM, Newble DI, eds. *International Handbook of Research in Medical Education*. Vol 2, Dordrecht, The Netherlands: Kluwer Academic Publishers; 2002:647-72.
2. Jackson T, Draugalis J, Smith E, Zachry W, III, Slack M. Hierarchical structure evaluation of Bloom's Taxonomy and Marzano's New Taxonomy using Rasch analysis. *Am J Pharm Educ*. 2002;66:108S.
3. United States Medical Licensing Examination. 2003 USMLE Bulletin of Information. United States Medical Licensing Examination. Available at: http://www.usmle.org/bulletin/2003/Overview.htm. Accessed July 27, 2003.
4. Ryan GJ, Nykamp D. Use of cumulative exams at US schools of pharmacy. *Am J Pharm Educ*. 2000;64:409-412.
5. Griffith PL, Rose J, Ryan JM. Student-curriculum maps: Applying the Rasch model to curriculum and instruction. *J Res Educ*. 1992;2:13-22.
6. Masters GN, Adams R, Lokan J. Mapping student achievement. *Int J Educ Res*. 1990;21:595-610.
7. Downing SM. Test item development: Validity evidence from quality assurance procedures. *Appl Meas Educ*. 1997;10:61-82.
8. Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. New York, NY: American Council on Education and Macmillan; 1989:13-104.
9. American Educational Research Association. American Psychological Association, National Council on Measurement in Education. *Validity. Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association; 1999:9-24.
10. Pray WS, Popovich NG. The development of a standardized competency examination for doctor of pharmacy student. *Am J Pharm Educ*. 1985;49:1-9.
11. Smythe MA, Slaughter RL, Sudekum MJ. Development of a comprehensive examination for baccalaureate pharmacy students prior to experiential teaching. *J Pharm Teach*. 1992;3:53-62.
12. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Adv Health Sci Educ*. 2002;7:235-241.
13. Smith EV, Jr. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *J Appl Meas*. 2001;2:281-311.
14. Jackson TR, Draugalis JR, Slack MK, Zachry WM, III, D'Agostino J. Validation of authentic performance assessment: A process suited for Rasch modeling. *Am J Pharm Educ*. 2002;66:233-242.
15. Draugalis JR, Slack MK, Sauer KA, Haber SL, Vaillancourt RR. Creation and implementation of a learning outcomes document for a Doctor of Pharmacy curriculum. *Am J Pharm Educ*. 2002;66:253-260.
16. American Council on Pharmaceutical Education. Accreditation Standards and Guidelines for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree. Chicago, Ill: The American Council on Pharmaceutical Education Inc; 1997.
17. American Association of Colleges of Pharmacy. Educational Outcomes. Alexandria, VA: Center for the Advancement of Pharmaceutical Education Advisory Panel on Educational Outcomes, AACP; 1998.
18. Marzano RJ. *Designing a New Taxonomy of Educational Objectives*. Thousand Oaks, Calif: Corwin Press, Inc; 2001.
19. Jackson TR, Draugalis JR, Smith EV, Zachry WM, III, Slack MK. Hierarchical Structure Evaluation of Bloom's Taxonomy and Marzano's New Taxonomy Using Rasch Analysis. *Am J Pharm Educ*. 2002;66:108S.
20. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15:309-334.
21. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Reprinted, University of Chicago Press; 1980.
22. Wright BD, Stone MH. *The Measurement Model. Best Test Design: Rasch Measurement*. Chicago Ill: Mesa Press; 1979:1-24.
23. Wolfe EW, Chiu CWT. Measuring pretest-posttest change with a Rasch rating scale model. *J Outcome Meas*. 1999;3:134-161.
24. Winsteps [computer program]. Version 3.45. Chicago, Ill: Mesa Press; 2003.
25. Wiggins G. Creating tests worth taking. *Educ Leadersh*. 1992;49:26-32.
26. Wright BD, Stone MH. Designing Tests. Best Test Design: Rasch Measurement. Chicago, Ill: Mesa Press; 1979:129-40.
27. Millman J, Greene J. The Specification and Development of Tests of Achievement and Ability. In: Linn RL, ed. *Educational Measurement*. 3rd edition. Phoenix, Az: The Oryx Press; 1993:335-66.
28. Linacre JM. Sample size and item calibration stability. *Rasch Measurement Transactions*. 1994;7:328.
29. United States Medical Licensing Examination. USMLE Clinical Skills Examination: a Fact Sheet. United States Medical Licensing Examination. May 31, 2002. Available at: http://www.usmle.org/news/newscse.htm. Accessed September 24, 2002.
30. Institute of Medicine (US) Committee on Quality Health Care in America. Setting performance standards and expectations for patient safety. In: Kohn LT, Corrigan JM, Donaldson MS, eds. *To Err is Human: Building a Safer Health System*. Washington, DC: National Academy Press; 2000:132-54.
31. Angoff WH, ed. Scales, norms, and equivalent scores. Washington DC: American Council on Education; 1971. Thorndike RL, ed. *Educational Measurement*.
32. Norcini JJ, Lipner RS, Langdon LO, Strecker CA. A comparison of three variations on a standard-setting method. *J Educ Meas*. 1987;24:56-64.
33. Institute for Objective Measurement Inc. Available at: http://www.rasch.org. Accessed September 24, 2002.