Nurse Educator

Nurse Educator
Vol. 34, No. 4, pp. 166-171
Copyright © 2009 Wolters Kluwer Health |
Lippincott Williams & Wilkins

How to Read and Really Use an Item Analysis

Thayer W. McGahee, PhD, RN
Julia Ball, PhD, RN

A frequent challenge for nursing faculty is to write a test that effectively evaluates learning and prepares students to be successful on the NCLEX-RN examination. Use of item analysis is an approach often used to provide an objective evaluation of examinations. Interpreting these analyses, however, can be frustrating. The authors provide an explanation of the various components of an item analysis, how to make an analysis useful for faculty, and how to use the components of an item analysis in revising tests.

Writing the perfect examination is an idealized and unrealistic goal for faculty, but one for which to continually strive. Faculty often spend hours writing and rewriting test questions each semester, only to find some new problem or issue with each version of the test. Test banks are frequently used in generating questions, but items are sometimes found to be poorly written,[1] so the frustration continues. Faculty often think they have written an excellent test, but without an objective item analysis it is really impossible to have any assurance of its evaluative competence. Using a computer-generated item analysis can be extremely useful to faculty in their constant endeavor to write the perfect test.

A well-written test serves to confirm that students are appropriately challenged, have a good grasp of the material that was taught, and are prepared to progress. Although many alternative forms of teaching and evaluation are available and frequently used in nursing curricula, the use of multiple-choice tests is the most common mode of objectively evaluating course work. This format is also used for initial licensure examinations, so creating good tests helps to prepare students to successfully take the NCLEX-RN licensing examination. Because licensure is the ultimate goal for a graduating nursing student, faculty have an obligation to prepare them well for this.

It has become common practice for nursing faculty to have an item analysis performed after administering an examination. A standard item analysis report yields a wealth of information, but to many faculty, the data are not meaningful, and therefore, the report is not used. When used appropriately, an item analysis can guide the faculty in revising and improving tests. There is, however, a dearth in the nursing literature on this subject.

Authors' Affiliation: Assistant Professor (Dr McGahee), Dean and Associate Professor (Dr Ball), College of Nursing, University of South Carolina Aiken, Aiken, South Carolina.
Corresponding Author: Dr McGahee, College of Nursing, University of South Carolina Aiken, 471 University Parkway, Aiken, SC 29801 (thayerm@usca.edu).

What Is an Item Analysis?

Slight variations exist in the statistics used in an item analysis, depending on the software used, but the general elements of the analysis are the same. Item analysis is a process of statistically examining both the test questions and the students' answers to assess the quality of the questions and the test as a whole.[2] The analysis assists in determining the extent to which individual test items contribute to the overall reliability, or internal consistency, of the test. The basic elements of an item analysis include measures of central tendency (mean, median, mode, standard deviation), correct group responses, response frequencies, nondistracters, point biserial, and a reliability coefficient. Looking at the item analysis in this order will provide a clear process to follow and enable faculty to systematically examine tests.

Measures of Central Tendency

Measures of central tendency are among the most basic of all the statistical results specified in an item analysis. The *mean* is simply the average of all individual student scores for a particular test or examination. The *median* is the number at which 50% of all scores on that test fall below. The *range* is the difference between the highest and lowest scores. *Standard deviation* is the measurement of variability. In other words, it is the measure of dispersion of student scores or how much on average scores vary around the mean. Although these statistics are easily understood by most, their interpretation may be somewhat skewed in a population of nursing students. For instance, in upper-level nursing courses, it is not expected that a percentage of students will fail the test or the course. In fact, the further along in the nursing curriculum the students are, the greater the expectation that students will pass. Therefore, while a mean of 80 might look attractive to a professor in a lower-level general education class, it may not be desired for a nursing test. If a mean such as this is obtained, it does not necessarily mean that the test is too hard or the students are not capable. Further investigation is required. It could be that there are 1 or 2 very low grades that are outliers and have skewed the mean.

Item Difficulty/Item Discrimination

In addition to measures of central tendency, other rather simple descriptive elements are in an item analysis. These include the correct group responses, response frequencies, and nondistracters (Table 1).

The correct group responses category is divided into 3 columns. One column gives the total percentage of students who answer an item correctly. This is a basic indicator of item difficulty. The greater the percentage of students answering a question correctly, generally, the easier that question is. By contrast, if a question has a zero percentage of students answering it correctly, that item does not contribute to distinguishing between individual differences among the students and is a question that needs to be revised. If a question has more than 50% incorrect responses, the faculty needs to examine the item and determine whether it needs to be revised or deleted or if there was a coding error.

The next column indicates the percentage of the upper third of the students answering an item correctly. The last of these 3 columns shows the percentage of the lower third of the students answering an item correctly. These last 2 columns of the correct group responses can be very useful. They tell you how the students making the highest grades and the lowest grades on this examination did on a particular question. This analysis is the first step in what is often called *item discrimination*. It would be expected that the upper third would do the best on all test items. However, if the reverse is true and students in the lower third do best on the item, it means that the question was not a good one or was worded poorly, thus misleading or discriminating against the students in the upper group. Any question that favors the lower third of the students and not the upper third needs to be revised. In addition, if an analysis shows low percentages in the upper third of the test takers selecting the correct answer, it warrants looking at those questions and determining if they need to be revised. A professor may have taught the content and the students just missed it, or it could be that the content explanation was not clear. In any event, if the upper third missed a test item, the professor needs to go back and reteach the material. The best test questions discriminate between those students who do well on the examination and those who do not.

The response frequencies are merely a tally of how many students responded to each of the possible answers to a particular question. It also usually indicates with an asterisk the correct response to that question. This is particularly useful when trying to understand why the majority of a class misses a question. For example, if no students answered an item correctly, it may be that the answer sheet was keyed incorrectly. Teacher error in keying is always a possibility. The response frequencies also help to illustrate which distracters are most challenging.

The nondistracters are the potential answers that none of the students chose as correct. When there are nondistracters, it means that the number of plausible answers is more limited than intended. In other words, if there are 4 potential answers in a multiple choice question, and 2 of those are nondistracters, the students were in essence answering a question with only 2 potential answers. Nondistracters are often too easy. All alternative answers should be plausible. It is important to remember that when writing questions for nursing examinations, they should mimic NCLEX-RN questions as closely as possible, so silly or obviously incorrect potential responses should not be used.

Point Biserial

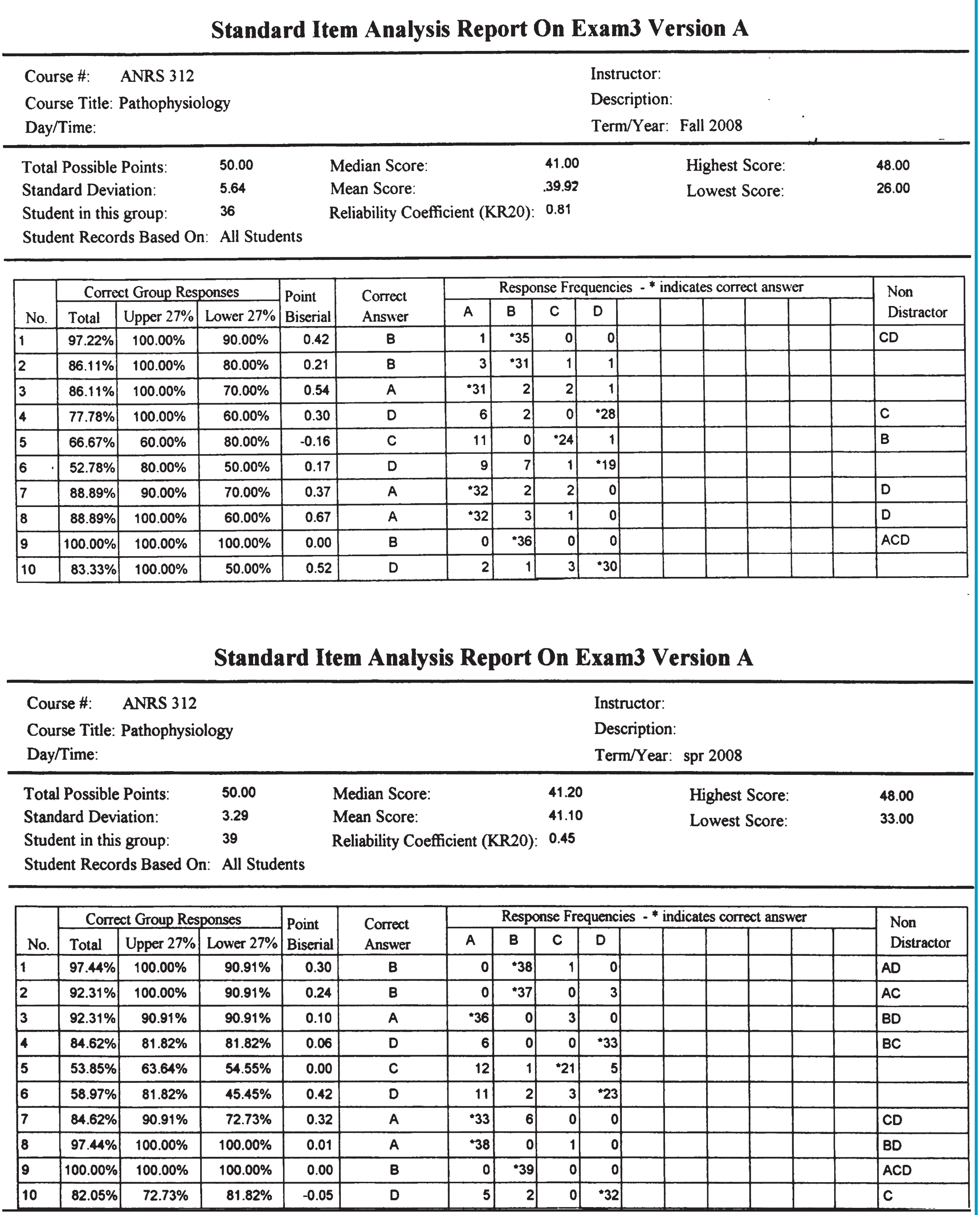
The point biserial is the second and more complete calculation of test item discrimination and is used to judge item quality. It tells how much predictive power a test item has and whether the students who would be expected to answer a question correctly are actually doing so.[2] The point biserial is a correlational calculation determined by the dichotomous variable of student responses to a particular test question (1 = right or 0 = wrong) and the continuous variable of their total score on the overall test. In other words, it is a correlation between item score and total score. This coefficient is an interaction between item discrimination and item difficulty. It is the measurement that illustrates how well an item separates, or differentiates, between those students who answer an item correctly or incorrectly, and have a high or low test score, respectively.[3] This number can range from −1 to +1. Very easy or very difficult test items will have little discrimination. Items of moderate difficulty (60%-80% answering correctly) generally are more discriminating.

The point biserial is designed to reflect the degree to which an item and the examination as a whole are measuring a single attribute topic and will be lower for examinations that measure a wider range of content. The higher the point biserial, the better that examination item is at discriminating among students on the basis of how well they really know the material.[4] A positive biserial indicates that those scoring higher on the test were more likely to answer that question correctly. If the students in the lower third answer an item correctly more frequently than the upper third of the students, the point biserial will have a negative value. This usually indicates that that test item is flawed and should be revised. A low value usually means that the question was too easy. There are no universal guidelines as to what point biserial value is most desirable on a nursing examination, but there are common ranges considered to be acceptable. As a general rule, anything below 0.20 is considered a poor question and in need of revision; items with a value between 0.20 and 0.30 are considered fair and could be improved upon, and items between 0.40 and 0.70 are considered good.[2,4,5] However, each question should always be evaluated in terms of the purpose of the test and the purpose of the individual question. For example, there may be a question that is so critical to the knowledge base of the students that the professor desires and expects 100% of the students to answer it correctly. In that case, a point biserial of 0 may be the goal.

Reliability Coefficient

The overall reliability of an examination is analyzed using a reliability coefficient. It may be reported as a Cronbach α

Table 1. ParSCORE Analyses for two 50-Item Examinations



Standard Item Analysis Report On Exam3 Version A

Course #: ANRS 312
Course Title: Pathophysiology
Day/Time:

Instructor:
Description:
Term/Year: Fall 2008

Total Possible Points:	50.00	Median Score:	41.00	Highest Score:	48.00
Standard Deviation:	5.64	Mean Score:	39.92	Lowest Score:	26.00
Student in this group:	36	Reliability Coefficient (KR20):	0.81		
Student Records Based On:	All Students				

	Correct Group Responses			Point	Correct	Response Frequencies - * indicates correct answer										Non
No.	Total	Upper 27%	Lower 27%	Biserial	Answer	A	B	C	D							Distractor
1	97.22%	100.00%	90.00%	0.42	B	1	*35	0	0							CD
2	86.11%	100.00%	80.00%	0.21	B	3	*31	1	1							
3	86.11%	100.00%	70.00%	0.54	A	*31	2	2	1							
4	77.78%	100.00%	60.00%	0.30	D	6	2	0	*28							C
5	66.67%	60.00%	80.00%	-0.16	C	11	0	*24	1							B
6	52.78%	80.00%	50.00%	0.17	D	9	7	1	*19							
7	88.89%	90.00%	70.00%	0.37	A	*32	2	2	0							D
8	88.89%	100.00%	60.00%	0.67	A	*32	3	1	0							D
9	100.00%	100.00%	100.00%	0.00	B	0	*36	0	0							ACD
10	83.33%	100.00%	50.00%	0.52	D	2	1	3	*30							

Standard Item Analysis Report On Exam3 Version A

Course #: ANRS 312
Course Title: Pathophysiology
Day/Time:

Instructor:
Description:
Term/Year: spr 2008

Total Possible Points:	50.00	Median Score:	41.20	Highest Score:	48.00
Standard Deviation:	3.29	Mean Score:	41.10	Lowest Score:	33.00
Student in this group:	39	Reliability Coefficient (KR20):	0.45		
Student Records Based On:	All Students				

	Correct Group Responses			Point	Correct	Response Frequencies - * indicates correct answer										Non
No.	Total	Upper 27%	Lower 27%	Biserial	Answer	A	B	C	D							Distractor
1	97.44%	100.00%	90.91%	0.30	B	0	*38	1	0							AD
2	92.31%	100.00%	90.91%	0.24	B	0	*37	0	3							AC
3	92.31%	90.91%	90.91%	0.10	A	*36	0	3	0							BD
4	84.62%	81.82%	81.82%	0.06	D	6	0	0	*33							BC
5	53.85%	63.64%	54.55%	0.00	C	12	1	*21	5							
6	58.97%	81.82%	45.45%	0.42	D	11	2	3	*23							
7	84.62%	90.91%	72.73%	0.32	A	*33	6	0	0							CD
8	97.44%	100.00%	100.00%	0.01	A	*38	0	1	0							BD
9	100.00%	100.00%	100.00%	0.00	B	0	*39	0	0							ACD
10	82.05%	72.73%	81.82%	-0.05	D	5	2	0	*32							C

ParSCORE, reproduced by permission from Scantron Corporation. Copyright© Scantron Corporation. All rights reserved. Scantron and ParSCORE are registered trademarks of Scantron Corporation.

or a Kuder-Richardson Formula 20 (KR-20) coefficient. This is a measure of the stability or consistency among the test scores or the internal consistency of a test.[4] The higher the reliability coefficient, the more likely a test will produce consistent scores when administered to similar groups. It is designed to measure how well a test measures a single cognitive factor. A low reliability coefficient may be reflected when a test covers multiple topics. The KR-20 is used for tests that have right and wrong answers. Cronbach α can be used for instruments that have right, wrong, and no right-wrong answer, such as in an attitude survey. For this reason, the KR-20 is most often used in education.

The KR-20 index ranges from 0 to 1, and reflects 4 different things: (1) the total number of test questions, (2) the proportion of the responses to an item that are correct, (3) the proportion of responses to an item that are incorrect, and (4) the variance for that set of scores.[6] Low reliability may mean that the test questions are unrelated to each other in terms of who answered them correctly and that the test scores reflect peculiarities of the test items more than the students' knowledge of the subject. The most common cause of a low reliability score is that the questions are too easy. Other reasons for a low reliability include an excessive number of very difficult items, unclear or poorly written items that do not discriminate, or test items that do not test a unified body of content.[7] A high reliability coefficient indicates that the individual questions on a test tended to pull together to measure 1 topic, and the students who did well overall were likely to answer each question correctly.

Practically speaking, a reliability coefficient of greater than 0.50 can be considered a good coefficient for a nursing examination because most nursing examinations cover multiple concepts and topics. Even if a test does cover multiple areas of content, such as cardiovascular and respiratory systems, the central construct for most examinations is still nursing knowledge, so the KR-20 is useful in analyzing the reliability of the test items relative to this construct. There are several ways to improve test reliability: administer longer tests, have a more heterogeneous group of students, or attempt to change the questions and thus the item difficulty to where 70% to 80% of the students answer test items correctly. However, none of these may be a viable option in nursing courses.

Examples of Test Analyses

Table 1 is an example of 2 different item analysis reports for 50-item examinations. Only the first 10 items are used for illustration. The first example has a high KR-20 of 0.81, indicating strong internal consistency. The mean of this examination is 79.84. The passing grade for this particular course is 80. There was 1 outlier of a grade of 52, which brought the mean down. Point biserials indicate that items 1 to 4 are good questions that discriminate well between the upper third of those scoring on this examination and the lower third. Item 5 is an example of the lower third scoring better than the upper, thus rendering a negative biserial. When this item was examined, it was found to be a knowledge-level question, so it is possible that the upper third of the class overanalyzed the question and the lower third did better because they simply took it at face value. Because the entire class answered item 9 correctly, the biserial is 0. The second example in Table 1 has a lower KR-20, but a higher mean. There is also more variability in the correct group responses.

Table 2 is an example of an item analysis of a 10-item quiz. As would be expected, most of the upper third of the students did very well on this quiz. Eight of the point biserials are above 0.20, indicating good discrimination between the upper and lower thirds of the class.

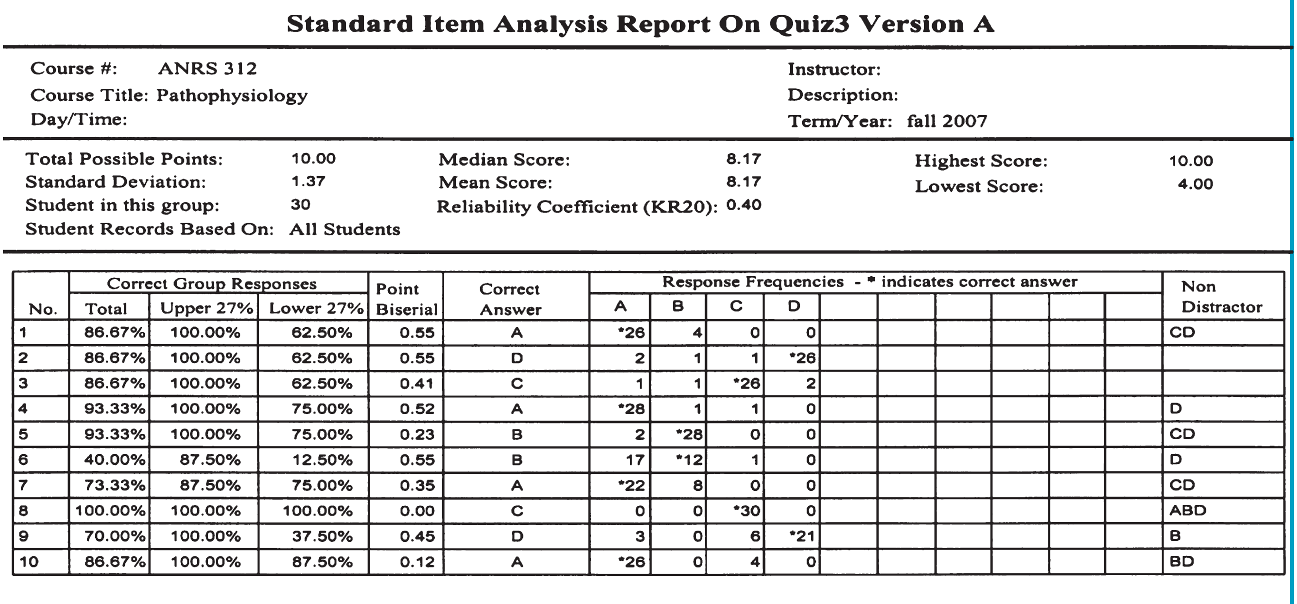
Now That I Can Read and Interpret an Item Analysis, So What?

The general interpretation of an item analysis can be more difficult when used with nursing students in the upper-level courses of the curriculum. The typical normal distribution of grades in a bell curve that might be expected in a freshman-level course should not be seen in this population of students. Nursing students in a baccalaureate nursing program should show a positive skewed distribution because they have already completed approximately 3 or 4 semesters of general education and science courses before they are accepted into a nursing program. The nursing major is a rigorous one, and the criteria for acceptance into nursing programs are more stringent than those of many other majors. There is nothing "average" about nursing students. This makes interpreting and using an item analysis even more complex because the basic rules of interpretation may not be valid for this population.

In nursing schools, everything has to be examined in context. A faculty member may administer an examination and have an item analysis that indicates that the test is an excellent one. It may have a reliability coefficient of 0.85, but a mean score of 76. If the cutoff for passing at a particular school is 80 and the mean score on an examination is 76, it means that the average score is not a passing one. This may be an indicator that the material for this examination was not grasped well by the students as a whole and needs to be reinforced or taught in a different way. It may also mean that the students simply did not prepare adequately. A good way to distinguish between these 2 causes is to administer the same or very similar test each semester and compare the analyses over time. Item analysis interpretations are usually normed with typical grading scales, so each nursing school, each course, and each examination must be considered when deciding what is acceptable on the item analysis and what indicates a need for a change in a particular examination. Caution should always be used when interpreting statistics based on a small sample size because the results may simply be random chance.

Evaluating examinations is difficult. Faculty members tend to like questions they write and want to think it is strictly the fault of the students if they miss questions. An item analysis is an objective measure to scrutinize examinations more critically and determine which questions really do need to be revised, what material needs to be revisited, and which questions need to be kept. If an item analysis indicates a poor question, it does not mean that the question must be discarded and everyone given credit for the item on that examination. It is most useful to look at the item analysis along with the test blueprint to see the

Table 2. ParSCORE Analysis for a 10-Item Quiz



Standard Item Analysis Report On Quiz3 Version A

Course #: ANRS 312
Course Title: Pathophysiology
Day/Time:

Instructor:
Description:
Term/Year: fall 2007

Total Possible Points:	10.00	Median Score:	8.17	Highest Score:	10.00
Standard Deviation:	1.37	Mean Score:	8.17	Lowest Score:	4.00
Student in this group:	30	Reliability Coefficient (KR20):	0.40		
Student Records Based On:	All Students				

	Correct Group Responses			Point	Correct	Response Frequencies - * indicates correct answer										Non
No.	Total	Upper 27%	Lower 27%	Biserial	Answer	A	B	C	D							Distractor
1	86.67%	100.00%	62.50%	0.55	A	*26	4	0	0							CD
2	86.67%	100.00%	62.50%	0.55	D	2	1	1	*26							
3	86.67%	100.00%	62.50%	0.41	C	1	1	*26	2							
4	93.33%	100.00%	75.00%	0.52	A	*28	1	1	0							D
5	93.33%	100.00%	75.00%	0.23	B	2	*28	0	0							CD
6	40.00%	87.50%	12.50%	0.55	B	17	*12	1	0							D
7	73.33%	87.50%	75.00%	0.35	A	*22	8	0	0							CD
8	100.00%	100.00%	100.00%	0.00	C	0	0	*30	0							ABD
9	70.00%	100.00%	37.50%	0.45	D	3	0	6	*21							B
10	86.67%	100.00%	87.50%	0.12	A	*26	0	4	0							BD

ParSCORE, reproduced by permission from Scantron Corporation. Copyright© Scantron Corporation. All rights reserved. Scantron and ParSCORE are registered trademarks of Scantron Corporation.

whole picture. Looking at the whole picture is helpful to see where to focus to develop new questions. It is also helpful to have colleagues look at tests and give feedback. An objective outsider will be able to discern awkward wording or poorly written questions.

Statistics from an item analysis are useful in understanding student performance on that examination, but the purpose of the examination should always remain paramount. There are both theoretical/conceptual and practical reasons for writing examination items and designing the test as a whole.[8] Results of an item analysis should always be carefully examined and not used as the sole determinant for rewriting or revising a test. Data from the analyses are always going to be influenced by the number of students taking the examination, the type of students, the variances in teaching, and the inevitable errors attributable to chance. Table 3 gives a brief summary of suggestions of when to revise and when not to revise questions.

When using a software package to do item analysis, it is possible to add results of a current examination to prior examinations to increase the sample size. This is helpful only if examination questions have not been changed. If the questions have been changed, it is helpful to compare the analyses to determine if the changes made the examination better.

Once a faculty member has learned how to read and interpret an item analysis, the analysis can be very helpful both in refining tests and also in indicating what may need strengthening or deleting in the teaching of the course. The analysis helps to find flaws or errors in a test so it can be adjusted before grades are posted. For instance, it may indicate that there are 2 right answers and both should be accepted or that a question was keyed incorrectly and the tests must be rescored. The analysis is also useful in determining which questions are too difficult or too easy so that those questions can be reworded or revised to be more appropriately challenging. When questions are found to have high levels of difficulty, it may be necessary for the material to be stressed more carefully in class or more fully explained. Identifying the common misconceptions

Table 3. What to Do With Test Questions

Upper Third	Lower Third	Biserial	Nondistracters Present	Action
Correct	Incorrect	+	−	None
Correct	Incorrect	+	+	Revise nondistracters
Incorrect	Correct	−	−	Revise test item
Incorrect	Correct	−	+	Revise test item and nondistracters
Incorrect	Incorrect	+ or −	+ or −	Reteach content

of the students by looking at the frequencies of the distracters also helps to determine material that needs to be further clarified in class.

Summary

Writing the perfect examination is a lifelong challenge, but the goal should be continual improvement. Interpreting examination results by use of an item analysis yields a wealth of information that is useful in both improving test items and in improving teaching. Because one of the goals of every nursing school is to have its students pass the NCLEX-RN examination on the first try, the time and effort given to interpreting examinations and using this information are invaluable.

References

1. Masters JC, Hulsmeyer BS, Pike MaryE, Leichty K, Tiller MT, Verst AL. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *J Nurs Educ*. 2001;40(1):25-32.
2. Matlock-Hetzel S. Basic concepts in item and test analysis. Paper presented at Annual Meeting of the Southwest Educational Research Association; January 23-25, 1997; Austin, Texas.
3. McGill exam results. Available at http://www.mcgill.ca/ncs/products/exams/. Accessed January 14, 2009.
4. Oermann MH, Gaberson KB. *Evaluation and Testing in Nursing Education*. 2nd ed. New York, NY: Springer Publishing Co; 2006:173-176.
5. Introduction to Item Analysis. Scoring Office: Academic Technology Services. Available at https://www.msu.edu/dept/soweb/itanhand.html. Accessed January 14, 2009.
6. Bodner GM. Statistical analysis of multiple-choice exams. *J Chem Educ*. 1980;57(3):188-190.
7. Kehoe J. Basic item analysis for multiple-choice tests. Practical assessment, research & evaluation. 1995;4(10). Available at http://PAREonline.net/getvn.asp?v=4&n=10. Accessed January 14, 2009.
8. Brown JD. Questions and answers about language testing statistics. *Shiken: JALT Testing & Evaluation SIG Newsletter*. 2001;5(3):12-15.

Americans Are Not Getting Enough Vitamin D

Studies released from the University of Colorado, Denver School of Medicine show that currently 3 out of 4 Americans have insufficient levels of vitamin D, often known as the "sunshine vitamin." The ideal level of this vitamin is between 30 and 40 nanograms per milliliter. The percentage of Americans with less than ten nanograms per milliliter has tripled in the last twenty years.

Dr. Adit Ginde notes that recent studies have uncovered numerous roles for vitamin D in addition to the well known role of this vitamin in bone health and the prevention of rickets. Many types of cells show receptors specific for vitamin D and over 100 different genes are known to be regulated by vitamin D. Vitamin D has also been shown to have a significant role in regulating immune system activity as well as preventing both cancer, and cardiovascular disease.

Obtaining enough vitamin D is simple. A person need only spend ten minutes in the sun with legs and arms exposed to trigger development of adequate amounts of this vitamin. Dr. Ginde notes however that people have changed sun-exposure habits in response to vigorous campaigns related to the prevention of skin cancers. The diligent use of sunscreen, clothing and hats unfortunately slows the body's production of vitamin D. Also, Americans have become less active over the last twenty years and spend less time out of doors.

Ginde notes that most Americans could actually use more vitamin D. Supplements are beneficial, as are small amounts of sun exposure. Consuming foods that are high in Vitamin D such as salmon, sardines, mackerel, and tuna along with dairy products will increase dietary intake of this vitamin. As the significance of vitamin D becomes better known, people need to attend to the methods to increase its levels.

We emphasize the dangers of sun exposure to our students and patients. We need to temper this information with responsible explanations of the need for vitamin D and methods to increase the amount of this vitamin to healthy levels.

Source: Stern, A. March 23, 2009. Americans need more Vitamin D: researchers. Reuters. Available at: http://www.reuters.com/article/healthNews/idUSTRE52M6M120090323. Accessed March 26, 2009.