An Item Analysis and a Reliability Estimate of a Classroom Kinesiology Achievement Test

Kyle Perkins
(Retired Professor)
Florida International University

Eva Frank
Lebanon Valley College

Publication Date: October 10, 2018

An Item Analysis and a Reliability Estimate of a Classroom Kinesiology Achievement Test

Abstract

This paper presents item-analysis data to illustrate how to identify a set of internally consistent test items that differentiate or discriminate among examinees who are highly proficient and nonproficient on the construct of interest. Suggestions for analyzing the quality of test items are offered on the construct of interest. Suggestions for analyzing the quality of test items are offered as well as a pedagogical approach to augment the time-on-task on higher-level cognitive tasks.

An Item Analysis and a Reliability Estimate of a Classroom Kinesiology Achievement Test

Introduction

The purpose of this paper is to present item-analysis data and a reliability estimate of test scores from an undergraduate classroom kinesiology achievement test and to illustrate how item-analysis data can be used to improve teaching and learning, and to improve the quality of classroom achievement test items. Item-analysis data can be used to provide a basis for remedial work, to identify areas that need more extensive attention, and to suggest curricular revisions, or shifts in teaching emphasis (Gronlund and Linn, 1990). Item discrimination indices are used to identify items that should be retained and items that need to be eliminated or drastically revised. Item-analysis data also identify those items that form an internally consistent scale as estimated by coefficient alpha, or by the Kuder-Richardson Formula 20 (Spector, 2013).

For this research, we retained items with a proportion of correct answers (*p*-values) from 0.33 to 0.67, with an item-total or item-remainder correlation greater than 0.30, and with a minimum estimate of reliability of 0.80.

Item Difficulty

Researchers such as Tuckman (1978) and Henning (1987) propose that items with a proportion of correct answers that is less than 0.33 or that is greater than 0.67 be rejected. However, there

are at least three reasons why "too easy" items and "too difficult" items may need to be retained. First, many test developers begin each test section with a few very easy items in order to overcome examinees' psychological inertia. Usually, these "warm-up" items are not scored (Henning, 1987). Second, Popham (1978) and Gronlund and Linn (1990) discuss the necessity of including specific content in a test. If all items at the extreme tails of the difficulty continuum are rejected, this may have serious consequences of the test's being insensitive to the instructional objectives being tested. These shortcomings will be reflected in content-related evidence in a validation study.

Gronlund and Linn (1990) make a succinct argument for including easy and difficult items in a mastery test:

> "The difficulty of the test item in a criterion-referenced test is determined by the nature of the specific learning tasks to be measured. If the learning tasks are easy, the test items should be easy. If the learning tasks are of moderate difficulty, the test items should be of moderate difficulty. No attempts should be made to modify item difficulty or to eliminate easy items from the test in order to obtain a range of test scores. On a criterion-referenced test we would expect all, or nearly all, pupils to obtain high scores when the instruction has been effective" (p. 131).

We maintained the 0.33 to 0.67 range of item difficulty for our research, because we have no idea of what item difficulty an easy item should have nor of what item difficulty a difficult item should have.

A third reason for relaxing the 0.33 to 0.67 proportion constraint for retaining items is "shaping the information curve. By systematically sampling items, it is possible to create a test that is more sensitive or discriminating at a given cut-off score or scores" (Henning, 1987, p. 50).

Item Discrimination

Item difficulty is an important consideration in terms of retaining or rejecting a given test item, but it does not provide sufficient information alone. A researcher must also consider item discriminability, that is, how well an item differentiates or discriminates among examinees who are highly proficient and nonproficient on the construct of interest (Henning, 1987).

For this research, we used the item-remainder coefficient which is the correlation of each item with the combination (sum or average) of all the remaining items, not counting that one.

> "If there are 10 items, then the item-remainder coefficient for the first item will be the correlation of item 1 with the combination of items 2 through 10. The item-remainder for the second item will be the correlation of item 2 with the combination of items 1 plus 3 through 10. The larger the item-remainder, the more the item in question relates to the remaining items" (Spector, 2013, pp. 178-9).

We chose the item-remainder coefficient as a discrimination index rather than the traditional point biserial correlation. The point biserial correlation is a correlation between item responses and total scores for a given test which can result in an inflated correlation coefficient, because part of the magnitude of the coefficient is due to the presence of the item in the total score itself. Kingston and Kramer (2013) report that point biserial correlations have an expected inflation of 0.32 on a 10-item test and an expected inflation of 0.10 on a 100-item test. We followed Kingston and Kramer's recommendation of a selection/retention criterion of an item-total correlation, or, in the case of the present research, an item-remainder coefficient greater than 0.30.

Estimate of Internal Consistency Reliability of the Test Scores

"Reliability is a matter of degree and is usually expressed by indices ranging from 0.00 to 1.00. Reliability can only be estimated and not truly calculated" (Genesee & Upshur, 1996, p. 62). Gronlund and Linn (1990) remind us that

> "reliability refers to the *results* obtained with an evaluation instrument and not to the instrument self. Any particular instrument may have a number of different reliabilities, depending on the group involved and the situation in which it is used. Thus, it is more appropriate to speak of the reliability of 'test scores' or of the 'measurement' than of the 'test' or the 'instrument'" (p. 78).

Coefficient alpha is commonly used as an estimate of internal consistency reliability. In this research, we used the Kuder-Richardson Formula 20, which is a special case of coefficient alpha and is appropriate for dichotomously scored items. Kuder-Richardson estimates of reliability provide an indication of the degree to which the items in the test measure similar characteristics. We set 0.80 as the minimum alpha for our estimate of internal consistency reliability of undergraduate classroom kinesiology achievement test scores, following the recommendation of Nunnally (1978) and of Lance, Butts, and Michels (2006).

Method

Subjects

Data collection began during the fall 2012 semester and continued through spring 2014. Data were collected from two sections of an undergraduate kinesiology class during fall 2012 and spring 2013, one section in summer 2013, and two sections in spring 2014. Each class was taught once per week for 160 minutes. One hundred eight students participated in the study. Anatomy is a prerequisite course for kinesiology.

Instrumentation

A 42-item test assessing students' knowledge regarding the origin, insertion, and primary moving action of the shoulder nomenclature, arthokinematics during shoulder movements, and other anatomical shoulder joint characteristics was prepared for the study. The test consisted of 22 multiple-choice questions (items 1-22), eight true-false questions (items 23-30), and 12 matching questions (items 31-42). The test was designed to provide a measure of performance that should be interpretable in terms of a clearly defined and delimited domain of learning tasks.

A multiple-choice item presents a problem and a set of alternative solutions. The "correct" answer is the best or correct solution to the problem. The incorrect distracters are meant to distract the uninformed student from the correct answer. Test developers have used multiple-choice items to measure a variety of learning outcomes at the knowledge and understanding levels.

True-false items are typically used to measure the identification of the correctness of statements of fact, definitions of terms, and statements of principles. Typically, they are not useful beyond the knowledge area.

Matching questions are limited to measuring factual information based on simple associations (Gronlund & Linn, 1990).

Each item was coded for Bloom's cognitive process dimension and Bloom's knowledge dimension. The cognitive process dimensions included analyze, order; analyze, explain; apply, calculate; apply, classify; analyze, differentiate; evaluate, conclude; remember, list; remember, describe; understand, predict; and understand, interpret. The knowledge dimension included conceptual, factual, and procedural.

Procedure

Before the achievement test was administered, several opportunities to learn were made available. A 105-minute lecture was delivered on the origin, insertion, and primary moving action of the shoulder musculature, the arthokinematics of the shoulder during movement, and other anatomical shoulder joint characteristics. Homework was assigned, to be completed before the test was administered. The students were directed to complete note cards on the topic. One week later, another 105-minute lecture followed on the same topic. The test was administered one week after the second lecture.

The actual instruction and the active learning experiences based on the lectures focused on all levels of cognition. The nature of instruction took into consideration what the literature discusses about millennial students who are known to be high-achieving, team oriented, and kinesthetic learners (Montenery, Walker, Sorensen, Thompson, Kirklin, White, & Ross, 2013). These students develop critical thinking skills most effectively through active learning experiences (Montenery et al., 2013).

Although some of the instruction was lecture based, the instructor of the course attempted to introduce content through active methods. Active learning includes interacting with others, giving a presentation, developing and participating in real experiences (e. g., biomechanical analysis of a peer). The homework assigned to the students prior to attending class was prework, which required students to engage in the material prior to attending the lecture. Holding students accountable to complete the homework opened up an opportunity for the instructor to become a facilitator of learning and for the student to become more actively engaged in the learning environment (Alharbi, 2015; Jensen, Kummer, & Godoy, 2015), especially for the content from the homework. For example, students were required to learn about the scapulothoracic rhythm, which is the biomechanical movement of the shoulder blade, prior to attending class by dissecting the concept on a notecard. Once students attended class, the instructor required students to work in pairs and evaluate each other's scapulothoracic rhythm by palpating the scapula (shoulder blade) as it moves when the arm moves through the ranges of motion of abduction. This student-centered approach provided the instructor with an opportunity to increase interaction between faculty and students as well as student and content (Stone, 2012).

Results

Table 1

Item Difficulty and Item Discrimination Indices for the Retained Items

Item Number	Item Difficulty	Item Delta	Item-remainder Coefficient	Item Type
1	0.41	12.09	0.35	M-C
3	0.44	12.40	0.39	M-C
8	0.34	11.35	0.36	M-C

22	0.59	13.91	0.34	M-C
31	0.50	13.00	0.49	Matching
32	0.58	13.81	0.50	Matching
33	0.59	13.91	0.38	Matching
34	0.49	12.90	0.54	Matching
36	0.57	13.71	0.45	Matching
37	0.63	14.33	0.56	Matching
38	0.52	13.20	0.55	Matching
39	0.55	13.25	0.52	Matching
40	0.56	13.60	0.54	Matching
41	0.63	14.33	0.60	Matching
42	0.39	11.88	0.34	Matching

KR-20 = 0.85

Table 1 presents the item difficulty, delta, and item discrimination indices, as well as the item type for the 15 retained items. Item difficulty (percent correct) is an ordinal measure. We used the Excel program function, NORM.S.INV, to transform the item difficulty values to *z*-scores that correspond to the percent of a normal distribution. Because *z*-scores can sometimes have negative values, we linearly transformed the *z*-scores to a delta metric: delta = 13 + 4 x *z* (Kingston & Kramer, 2013). The point here is that item difficulties are ordinal measures; deltas are equal interval measures.

The 15 retained items (36 percent of the total items) had a KR-20 internal consistency estimate of 0.85 which exceeded our reliability criterion of 0.80. Eighteen percent of the multiple-choice items were retained, none of the true-false items were retained, and 92 percent of the matching items were retained. “Retained” in this context means that the items met the criteria that we set for item difficulty and item-remainder discrimination.

The 15 retained items assessed the primary purpose of the shoulder blade, a shoulder complex attachment, degrees of freedom at the sternoclavicular joint, lateral rotation of the glenohumeral joint, and the origin, insertion, and action of the serratus anterior, triceps, upper

trapezius, posterior deltoid, latissimus, pectoralis major, teres minor, anterior deltoid, middle deltoid, biceps brachii, and middle trapezius

For the retained items, we found a moderate relationship between the deltas and the item-remainder coefficients, $r = 0.58$, $p < .05$, df 13, two-tailed test.

Table 2

Item Difficulty and Item Discrimination Indices for the Non-Retained Items

Item Number	Item Difficulty	Item Delta	Item-Remainder Coefficient	Item Type
2	0.22	9.91	0.06	M-C
4	0.50	13.00	0.16	M-C
5	0.10	7.87	0.14	M-C
6	0.44	12.40	0.22	M-C
7	0.21	9.77	0.03	M-C
9	0.26	10.43	0.12	M-C
10	0.54	13.40	0.29	M-C
11	0.31	11.02	0.12	M-C
12	0.56	13.60	-0.11	M-C
13	0.63	14.33	0.18	M-C
14	0.67	14.76	0.20	M-C
15	0.45	12.50	0.23	M-C
16	0.88	17.70	0.17	M-C
17	0.91	18.36	0.09	M-C
18	0.72	15.33	0.15	M-C
19	0.38	11.78	0.05	M-C
20	0.34	11.35	0.08	M-C
21	0.40	11.99	0.04	M-C
23	0.84	16.98	0.09	T-F
24	0.69	14.98	0.16	T-F
25	0.47	12.70	0.06	T-F
26	0.70	15.10	0.10	T-F
27	0.76	15.82	-0.02	T-F
28	0.31	11.02	0.27	T-F
29	0.41	12.09	0.14	T-F
30	0.83	16.83	0.17	T-F
35	0.38	11.78	0.29	Matching

Table 2 presents the item difficulty, delta, and item discrimination indices, as well as the item type for the 27 items (64 percent of the total items) that were not retained and will be the focus of revision and future research, which will go far beyond the scope of this paper. Eighteen of the 22 multiple-choice items were not retained, all eight true-false items were not retained, and only one matching item (eight percent of the 12) was not retained.

The 27 items assessed the shoulder complex, glenoid labrum, humeral head, scapulothoracic joint motion, scapula tilting, sternoclavicular joint protraction, sternoclavicular joint classification, sternoclavicular joint elevation, acromioclavicular joint, classification, scapular movement, scapulothoracic joint classification, sternoclavicular joint protraction, glenohumeral-scapula motion ratio, glenohumeral joint classification, humeral head size comparison, coracohumeral ligament's primary responsibility, biceps' contraction results, lateral rotation and extension effect on the glenohumeral joint, cause and effect relation of movement of the scapula, scapula depression effect, abduction effect on the humerus, passive translation effect at the glenohumeral joint, resting position of the glenohumeral joint, formation of the coracoacromial arch, coracoclavicular ligament components, location of the jugular notch, and the origin, insertion, and action of the rhomboids.

For the non-retained items, there was no relationship between the deltas and the item-remainder coefficients, $r = -0.01$.

Discussion and Future Research

What is truly surprising to us is the preponderance of matching items in Table 1, the complete absence of true-false items, and a scant number of multiple-choice items. The items listed in

Table 2, which have low indices of discrimination, may indicate that these items are measuring something different from the items listed in Table 1, i.e., the test may be multidimensional.

The next phase of our research will begin with an analysis of the quality of each item in Table 2, using the following questions from Gronlund and Linn (1990):

1. Is the item format appropriate for the learning outcome being measured?

2. Does the knowledge, understanding, or thinking skill called forth by the item match the specific learning outcome and subject-matter content being measured?

3. Is the point of the item clear?

4. Is the item free of excessive verbiage?

5. Is the item of appropriate difficulty?

6. Does the item have an answer that would be agreed upon by experts?

7. Is the item free from technical errors and irrelevant clues? These include (1) grammatical inconsistencies, (2) verbal associations, (3) specific determiners (i.e., words such as always and never), and (4) some mechanical features, such as correct statements tending to be longer than incorrect ones?

8. Is the item free from racial, ethnic, and sexual bias? (pp. 230-232).

We believe that a thorough analysis of the non-retained items, answering the first two questions shown above, will provide insight on the behavior of the items shown in Table 2.

Another focus of our future research will be content-related evidence for validation: Does the sample of test tasks represent the domain of tasks to be measured? Is the test content representative of the knowledge, skills, and abilities that the teacher has taught? According to Gronlund and Linn (1990), items that represent an area receiving little emphasis tend to have poor discriminating power.

When we replicate this study, we will prepare a two-way table of specifications to include subject matter content (the topics to be learned) and instructional objectives (the types of performance students will be expected to demonstrate (remember, understand, apply, analyze, evaluate, and create). An achievement test should represent the content area and the objectives that are intended to be assessed. We will also track the emphasis or time-on-task for each content area and each instructional objective and ensure that the achievement domain is in harmony with what was taught. It should be noted that this is not teaching to the test.

We are particularly concerned about items that are written at higher cognitive levels. One of the authors coded the cognitive level of each item in the test. Another researcher independently coded the items, and there was an 88 percent agreement between the two raters. A third party adjudicated the discrepancies. It is often difficult to differentiate remembering and understanding items, because the difference may depend upon nuances of language used during instruction. The items were coded RU for remembering or understanding, and the other code was for items at a higher cognitive level, i.e., application, analysis, evaluation, or creation.

Table 3

Cognitive Level of Retained Items

Item Number	Cognitive Level
1	
3	RU
8	RU
22	RU
31	RU
32	RU
33	RU
34	RU
36	RU
37	RU
38	RU

39	RU
40	RU
41	RU
42	RU

RU = Remembering or Understanding

Blank indicates a higher cognitive level

Table 4

Cognitive Level of Non-Retained Items

Item Number	Cognitive Level
2	
4	
5	
6	
7	
9	
10	RU
11	
12	RU
13	
14	RU
15	RU
16	
17	RU
18	
19	
20	
21	
23	
24	
25	
26	RU
27	
28	RU
29	RU
30	RU
35	RU

RU = Remembering or Understanding

Blank indicates a higher cognitive level

Table 3 presents the cognitive level of each of the retained items. Ninety-three percent of them were coded at the lowest cognitive level. Table 4 presents the cognitive level of each of the non-retained items. Seventeen (63 percent) of the non-retained items were coded at a higher cognitive level than remembering or understanding.

It may have been the case that the 17 items listed in Table 4, which were coded at a higher cognitive level than remembering and understanding, were more difficult than the items coded as remembering and understanding. To answer that question, we summed and averaged the deltas for the RU items and for the higher cognitive level items. The mean for the RU items was 13.94; the mean for the higher cognitive level items was 12.79. A Kruskal-Wallis test indicated a significant difference in the ranks, chi-square = 23.31, $p < .01$, df 1.

The difficulty of the items may have been one cause of the lower discriminability of the non-retained items. Our experience with this test and other research indicates that more time-on-task needs to be applied on higher cognitive skills that students are expected to demonstrate with test items requiring application, analysis, evaluation, and creation. Fremer (2013) hypothesized that items written at the lowest levels of cognition, i.e., remembering and understanding (Anderson & Krathwohl, 2000), likely would be more instructionally sensitive, while items written at a higher cognitive level would be less instructionally sensitive. Roid and Haladyna (1982) define instructional sensitivity as:

> "the tendency for test items to range in difficulty as a function of instruction. Items that do not detect differences from pretest to posttest should be reviewed to ascertain if (a)

instruction has been faulty or (b) the item is flawed in some way or simply inappropriate" (p. 218).

It may be the case that higher cognitive level test questions are not as instructionally sensitive as lower cognitive level questions. This is also an area for ongoing research. Based on the results presented in this paper, we offer a pedagogical recommendation that Bruner's (1960) spiral curriculum be employed to increase time-on-task on learning objectives that reflect higher cognitive level tasks. Bruner postulated that as a curriculum, i.e., a program of instruction such as a class, evolves and develops, it "should revisit the basic ideas repeatedly, building upon them until the student has grasped the full formal apparatus that goes with them" (p. 8). This cycling and recycling process is an example of what Bruner refers to as the spiral curriculum.

Conclusion

We have shown how item analysis, item difficulty and item discrimination indices, can be used to identify an internally consistent set of items that discriminate between proficient and nonproficient examinees. The results indicate that a second group of items may need serious analysis and possible revision and then pilot tested again. It may also be the case that the test that we analyzed may be multidimensional. That is an agenda item for another research project. We also introduced Bruner's spiral curriculum as one means of improving teaching, learning, and students' performance on an undergraduate classroom kinesiology achievement test.

References

Alharbi, A. H. (2015). A flipped learning approach using social media in health informatics education. *Creative Education, 6,* 1466-1475.

Anderson, L. W., & Krathwohl, D. R. (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Boston: Pearson.

Bruner, J. S. (1960). *The process of education.* Cambridge, MA: Harvard University Press.

Fremer, J. (2013, November). Debate on the use of instructional sensitivity information to select test items for state tests. Invited debate at the Instructional Sensitivity Conference, Lawrence, KS.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education.* Cambridge, UK: Cambridge University Press.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th edition). New York: Macmillan Publishing Company.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Rowley, MA: Newbury House Publishers.

Jensen, J. L., Kummer, T. A., & Godoy, P. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE-Life Sciences Education, 14,* 1-12.

Kingston, N. M., & Kramer, L. B. (2013). High-stakes test construction and test use. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods. Volume 1: Foundations* (pp. 189-205). New York: Oxford University Press.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The source of four commonly reported

cutoff criteria: What do they really say? *Organizational Research Methods, 9,* 202-220.

Montenery, S. M., Walker, M., Sorensen, E., Thompson, R., Kirklin, D., White, R., & Ross, O. (2013). Millennial generation student nurses' perception of the impact of multiple technologies on learning. *Nursing Education Perspectives, 34,* 405-409.

Nunnally, J. C. (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* New York: Academic Press.

Spector, P. E. (2013). Survey design and measure development. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods. Volume 1: Foundations* (pp. 170-188). New York: Oxford University Press.

Stone, D. B. (2012). Flip your classroom to increase active learning and student engagement. 28th Annual Conference on Distance Teaching and Learning, 1-5. Retrieved from http://www.uwex.edu/disted/conference/Resource_library/proceedings/56511_2012.pdf

Tuckman, B. W. (1978). *Conducting educational research* (2nd ed.). New York: Harcourt Brace Jovanovich.