International Journal of Psychological Studies; Vol. 7, No. 1; 2015
ISSN 1918-7211 E-ISSN 1918-722X
Published by Canadian Center of Science and Education

A Novel Method for Evaluating Examination Item Quality

Kenneth D. Royal[1] & Mari-Wells Hedgpeth[2]

[1] Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA

[2] Office of Educational Development and Research, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA

Correspondence: Kenneth D. Royal, Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, 1060 William Moore Dr., Raleigh, North Carolina 27695, USA. E-mail: kdroyal2@ncsu.edu

Received: December 21, 2014 Accepted: January 8, 2015 Online Published: February 17, 2015

doi:10.5539/ijps.v7n1p17 URL: http://dx.doi.org/10.5539/ijps.v7n1p17

Abstract

Background and aims: Poor quality examination items may result in invalid test scores that potentially misrepresent what a student actually knows about a given content area. Thus, if an examination consists largely of poor quality items it is plausible that an individual with minimal content knowledge could perform reasonably well and receive a score that erroneously inflates his or her measure of ability. This study builds on this premise by introducing a novel method for evaluating item quality and demonstrating its utility. Method: We sought to understand the extent to which medical school examination items were vulnerable to good test-taking skills and guessing strategies by administering an examination to a group of medical education professional staff. The extent to which persons with no formal medical training could perform above the odds of random guessing were used to identify zones in which items may be vulnerable to guessing strategies. Results: The performance of professional staff was able to provide excellent diagnostic information regarding which items may be particularly vulnerable to guessing strategies. Conclusions: The proposed methodology was demonstrated to be successful, thus we encourage other medical educators to adopt this model for evaluating item and examination quality in a variety of contexts.

Keywords: psychometrics, testing, validity, item quality, guessing, medical education

1. Introduction

It has been well documented in the research literature that poor quality examination items may result in invalid test scores that potentially misrepresent what a student actually knows about a given content area (Downing & Haladyna, 1997; Downing & Haladyna, 2004; Haladyna, Downing, & Rodriguez, 2002; Penfield, 2013). This is primarily because poor items often offer psychological cues within the wording of an item or its response options that may unduly benefit a test-taker or otherwise improves one's chance of answering the item correctly. If an examination consists largely of poor quality items it is plausible that an individual with minimal content knowledge could perform reasonably well and receive a score that erroneously inflates his or her measure of ability. The extent to which this score contamination is likely to become a problem could be based in part on one's test-taking skills.

Currently, most medical educators rely primarily on item difficulty estimates (p-values), discrimination coefficients, and distractor analyses to determine an item's quality. Although the psychometrics literature offers a number of additional techniques for discerning item quality, many of these methods are quite sophisticated and require advanced training in measurement and statistics rendering them inaccessible for most medical educators. The purpose of this study was to (1) introduce a simple method for evaluating item quality, and (2) use the methodology to examine how well a sample of educated people with no medical school training could perform on the easiest exam questions asked of second year medical students in an organ-based curriculum of the Gastrointestinal Systems (GI).

2. Method

2.1 Design

We sought to understand the extent to which medical school examination items were vulnerable to good test-taking skills and guessing strategies by administering an examination to a group of medical education professional staff. Each professional staff member was directed to do his or her best on the examination and presented an examination under similar conditions as actual medical students. Using Classical Test Theory (CTT) psychometric methods, item statistics were evaluated to determine which items possessed potential vulnerabilities and should be candidates for revision or removal from future examinations.

2.2 Sample

Professional staff from a large United States medical school's Office of Medical Education (OME) were invited to participate in a study that attempted to determine the extent to which various medical school examination items were vulnerable to good test-taking skills and guessing strategies. Staff members were employed in various departments under the OME umbrella, such as Student Affairs, Admissions, Finance, and Curriculum Support offices, etc. To qualify for inclusion in the study OME staff must hold at least a bachelor's degree and have no formal education or experiential training in the physical, life, health, or biomedical sciences. These criteria for inclusion were necessary in order to establish a reasonable baseline for an educated person with minimal medical content knowledge.

A total of 13 professional staff members volunteered to participate in the study. Eight participants were male and five were female. The age of participants ranged from 25-64 (M=40.38, SD=11.33) with a median age of 38. Twelve of the thirteen participants identified as White/Caucasian. The amount of time employed in medical education ranged from 1 year to 24 years (M=7.31, SD=7.10), with a median of 4 years' experience. With regard to highest degree earned, two participants held doctoral degrees in social science disciplines, six held master's degrees in social science and humanities fields, and five held bachelor's degrees primarily in the social and behavioral sciences and the humanities.

2.3 Course and Examination

A second year Gastrointestinal (GI) System course served as the setting for this study. The course ran approximately three and a half weeks in duration and was taught during the Fall semester block. The course enrolled approximately 180 second year medical students. Course instruction was provided by a team of GI faculty experts.

The examination for OME professional staff was prepared with three key factors in mind. First, we recognized that the easiest items for medical students were the strongest candidates for items with vulnerabilities to guessing, thus we opted to select only the easiest items for inclusion on the OME examination. Second, we realized it would be an unreasonable request to ask professional staff to voluntary complete an examination covering an excessive amount of items, particularly items that pertain to content with which they are most likely to be unfamiliar, so only a reasonable number of items could be used on the examination. Third, we decided to cap the examination at 60 items based on suggestions from the psychometric literature that indicates the size of standard errors associated with 60 items is quite small (Fisher, 2008).

The examination was developed by reviewing the item analysis statistics from the 2012-13 mid-term and final examinations. The 60 easiest items based on actual medical student performance yielded p-values (percent correct) of .89 or higher. All items were also evidenced to possess adequate discrimination estimates based on point biserial correlation calculations. The 2012-13 GI mid-term contained a total of 35 items with 27 (77.14%) having a difficulty of .89 or greater. The final exam administered as part of the 2012-13 GI course had a total of 70 items. The final exam was more challenging for students as only 33 (47.14%) items had a difficulty of .89 or greater.

2.4 Guessing Strategy

Accompanying each item on the OME examination was a follow-up question that asked staff members to rate the extent to which they relied on guessing strategies to answer the previous question. Rogers (1999) identified three types of guessing: random, cued, and informed. Random guessing refers to carelessly choosing a response to an item, cued guessing refers to providing a response based on one's test-wiseness, and informed guessing refers to making a guess based on partial knowledge. We would anticipate the odds of correctly answering an item would increase as one moves from random, to cued, and then to informed guessing tactics. Using Rogers (1999) framework for guessing, we asked test-takers to indicate whether they relied on random, cued, or informed guessing, or no guessing at all. Specifically, we provided the following item:

Please identify the strategy you used to answer the previous question from the options below:

1) I did not guess.

2) Informed guessing: I selected a particular answer based upon previous partial knowledge of the subject, or I was able to eliminate particular answer options based upon previous partial knowledge of the subject.

3) Cued guessing: I selected an answer based upon some sort of stimulus within the test such as wording cues, cues associated with item stems, choices among answer options, test-wiseness, etc.

4) Random guessing: I selected a particular answer by blindly choosing an answer.

2.5 Examination Procedures

We administered the examinations using a secure browser on students' medical school laptops. The testing software contained a database that houses historical information on items used from approximately 1999 to the present and we used that data to inform the construction of the exam used in this study. Having the ability to view each item's historical statistical performance was essential for selecting items that appeared to be potentially vulnerable to good test-taking skills.

To make the testing process as comparable as possible to a medical student's experience, all conditions were imitated as closely as possible with exception to the stakes associated with the results. More specifically, professional staff participants completed the exam in the same web-based testing system, through the exact same means that students routinely do. Participants were allowed two hours to complete the 60 item online exam. Medical students at this institution are typically allowed one minute and thirty seconds per item, but given this was the first time this sample of participants had interacted with the testing software we deemed two minutes per item acceptable. Further, because this exam was a "power" exam (Gulliksen, 1950; Royal, O'Neill, & Shirley-Akers, 2011) it was reasonable to assume staff performance would not dramatically improve even if they were allotted an unlimited amount of time.

Prior to being allowed to launch the online exam participants had to agree to the same ethical requirements of students, particularly confidentiality and non-disclosure statements prohibiting the writing down of any questions or responses from the exam, and sharing any information on the exam with anyone else. In addition, participants had to agree to complete the exam without assistance from any outside resources. All participants agreed to these terms. Permission to conduct this study was granted by the university's Institutional Review Board.

2.6 Analysis

Data analysis consisted of scoring OME professional staff performance on the examination, investigating various person and item statistics, and comparing staff performance to most recent students' performance on each of the same items. We also investigated the extent to which each staff member indicated s/he guessed on each item, as well as the type of guessing strategy that was purportedly used. Data analyses were conducted under the Classical Test Theory (CTT) framework using values produced by Winsteps measurement software (Linacre, 2014).

3. Results

At the overall test level, it is highly unlikely that an individual can do extremely well on a test of 40 or more well-written, psychometrically-sound items with little to no content knowledge. The probability of randomly guessing an item correctly would be 25% for items with four response options. Therefore, the likelihood of an examinee making a series of "lucky guesses" and producing a strong score is extremely unlikely, probably somewhat comparable to the odds of winning the lottery (Downing, 2003). Before analyzing the data, participant responses about their guessing behaviors were investigated to determine the extent to which guessing strategies were employed on the examination. The findings indicate staff participants relied almost entirely on some sort of guessing strategy (see Table 1).

Table 1. Type and frequency of guessing strategy used

Strategy Used	Count	Percent
Random Guess	460	59.126
Cued Guess	265	34.602
Informed Guess	46	5.913
Did not Guess	7	0.009
Total	778	100

Note. Two guessing responses were missing.

Percent Correct (p-value) scores ranged from 22% to 67%, with a mean score of 36.31% (SD = 11.68%). One individual appeared to be an outlier by answering 67% of items correctly. This individual answered 13 more items correct than the second highest scorer, thus scoring 22% points higher than the second highest performer. Because of the confidential nature of this examination, we did not follow-up with this individual to investigate possible reasons for such an unexpectedly high performance. Upon exclusion of this outlier, the collective group average performance dropped to 33.75% (SD = 7.48%).

The exam had an average of 4.23 response options per item. This corresponds to approximately 23.64% odds of success should one answer each item at random. By subtracting the random odds of success from the average group performance, one could essentially establish a baseline guessing error (BGE) estimate of how much testwiseness contamination appeared in this set of scores. Given the average score for the group was 36.30% with a median of 32%, we determined the exam contained approximately 8.36% to 12.36% BGE. It is important to note the BGE, as defined here, is simply an estimate of baseline guessing error that can be accounted for when assuming the participants have minimal content knowledge. In reality, guessing effects can never be perfectly accounted for in any statistical or measurement model, nor can one ever truly know to what extent persons with content knowledge may benefit from guessing on any given examination. In any instance, it was evident that staff performed just slightly above random odds of success on items that have been empirically evidenced to be very easy for medical students.

With regard to item statistics, an array of insightful information was obtained. Item difficulty estimates (p-values) were sorted from low to high and "zones" were established to indicate the extent to which the item may be potentially vulnerable to good test-taking skills. We elected to use the term "zones" because items that are particularly vulnerable to good test-taking skills are considered contaminated, thus zones nicely illustrate levels of potential contamination for items. We created four different zones according to a probability of success schema based on four to five response options per item. In particular, the beginning odds of success for an individual relying completely on random guessing would approximate 20-25% depending on whether the item offered four or five response options. Therefore, items with p-values less than 25% fell into the "safe zone". By eliminating one distractor on an item with four response options, an individual's odds of success improves to 33%. With this criteria in mind, we established a "low caution zone" for items with p-values ranging between 26 and 32% and a "caution zone" for items with p-values ranging between 33% and 49%. Finally, items with p-values greater than 50% fell into the "danger zone" as these items demonstrated psychometric evidence of being vulnerable to samples of examinees that have minimal medical content knowledge.

Using the aforementioned criteria, we determined 21 (35%) items fell into the "safe zone", 11 (18.33%) fell into the "low caution zone", 12 (20%) fell into the "caution zone", and 16 (26.67%) fell into the "danger zone". Considering each of the items that appeared on the OME professional staff examination was based on items that medical students had historically performed incredibly well on (p-value of .89 or higher), the results of this study indicate that approximately 65% of these items may have some elements of vulnerability that may potentially contaminate the measurement system, and potentially invalidate some of the scores.

4. Discussion

Professional staff indicated more than 99% of their responses on this examination were based on some type of guess. Given none of the participants had any formal educational or experiential training in any physical, life, health or biomedical sciences, one might reasonably assert the sample of individuals utilized in this study represents an educated group of individuals with a proximal baseline of Gastrointestinal medical content knowledge. Interpretation of results should take these important nuances into account.

It also stands to reason that all guessing strategies are not created equal. Conventional wisdom theorizes knowledge as a continuum and the farther one advances along this continuum from random to informed guessing, the greater one's odds of success should become. To analyze each item individually in light of the particular guessing strategy used could be useful, but with such a small sample size and a research literature that does not identify exactly how much one's odds of success might be expected to improve given each guessing type such an analysis was beyond the purview of this study.

Results of this study clearly indicate that some level of medical content knowledge is necessary to perform remotely well on the easiest of exam items for this course. Do the results of this study provide authentic evidence of student learning? There is certainly a great deal of evidence to support this possibility, but it should be noted that another distinct possibility is that these scores may be contaminated by the effects of "teaching to the test". It is possible that the medical students' scores are, at least in part, contaminated by the effects of exam items that may have been used as examples in class, or derived directly from lecture materials. Without more information regarding the source of each item's content and how the material was delivered it is impossible to know whether the scores are truly authentic measures of student learning, or a measure of student learning coupled with some instructional sensitivity (Popham, 2006) effects.

It is important to note that psychometric analyses of student performance and examination functioning are a must for medical school exams due to their moderate to high-stakes nature. Such analyses should be conducted after every major exam. New items should always be piloted and existing items with identifiable flaws should be revised before administering again. These are basic principles of good assessment. Unfortunately, many of the most useful and informative psychometric techniques for analyzing data are quite sophisticated and require significant training to perform. Similarly, traditional statistical analyses of items based on the Classical Test Theory (CTT) tradition have major limitations (e.g., scores are sample dependent to the test, raw scores do not take into account item difficulty, etc.) (Bond & Fox, 2007; Royal, 2010; Embretson & Reise, 2000; Royal, Gilliland, & Kernick, 2014). Most medical educators find themselves in a situation in which they are not sufficiently skilled in psychometrics to conduct the more rigorous and "gold standard" types of analyses, but they are sufficiently skilled to conduct elementary types of analyses that possess a number of significant limitations. The methodology presented in this paper offers medical educators a simple and accessible approach for understanding item and exam quality, while at the same time offering an additional lens for which to view data and investigate evidence for construct validity.

Limitations of this study obviously include the small sample. In particular, the statistical stability of the staff participants' scores may potentially introduce error that could cause some exam items to appear in different safe/caution zones. Further, the sample frame consisted of educated professional staff working in the field of medical education. It is unclear if working in medical education provided some inadvertent advantage over other potential participants, or if the performance of this particular sample may be representative of other samples of educated persons relying on various guessing strategies. Finally, the extent to which participants were motivated could also be a factor. Although participants were encouraged to perform their best, there were no stakes associated with the score results. Thus, it is difficult to know if participants would even select the same guesses on a second attempt at the exam.

5. Conclusion

This study attempted to understand the extent to which examination items in a GI course were vulnerable to good test-taking skills. Specifically, we constructed a 60 item exam based on the easiest items as determined by medical students' performance on mid-term and final exams. We then administered the exam to medical education professional staff. Results of the analysis revealed a great deal of useful information for the course instructor, such as which items were particularly free of potential guessing contaminants for persons with baseline medical knowledge, which items possessed guessing vulnerabilities for persons with minimal content knowledge, and the extent to which contamination may be a problem in the absence of medical knowledge. We believe studies of this kind will be of significant benefit for medical educators and encourage others to use this methodology as a model for evaluating item and exam quality.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Downing, S. M. (2003). Guessing on selected-response examinations. *Medical Education*, *37*, 670-671. http://dx.doi.org/10.1046/j.1365-2923.2003.01585.x

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*, 61-82. http://dx.doi.org/10.1207/s15324818ame1001_4

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*, 327-333. http://dx.doi.org/10.1046/j.1365-2923.2004.01777.x

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fisher, W. P. (2008). The cash value of reliability. *Rasch Measurement Transactions*, *22*(1), 1160-1163.

Gulliksen, H. (1950). Speed versus power tests. In *Theory of Mental Tests* (pp. 230-244). Hoboken, NJ: John Wiley & Sons Inc. http://dx.doi.org/10.1037/13240-017

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309-334. http://dx.doi.org/10.1207/s15324818ame1503_5

Linacre, J. M. (2014). *WINSTEPS® (Version 3.81.0)*. Computer Software. Beaverton, OR: Winsteps.com.

Penfield, R. (2013). Item analysis. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hanse, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology*, *Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 121-138). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14047-007

Popham, W. J. (2007). Instructional sensitivity on tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146-150,155. http://dx.doi.org/10.1177/003172170708900211

Rogers, H. J. (1999). Guessing in multiple-choice tests. In G. N. Masters, & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 235-243). Oxford, UK: Pergamon. http://dx.doi.org/10.1016/b978-008043348-6/50019-x

Royal, K. D. (2010). Making meaningful measurement in survey research: A demonstration of the utility of the Rasch model. *IR Applications*, *28*, 2-16.

Royal, K. D., O'Neill, T. O., & Shirley, K. (2011). *"Speed vs. Power": To what extent is speed a factor on the American Board of Family Medicine's certification examination?* Paper presented at the 2011 American Educational Research Association. New Orleans, LA.

Royal, K. D., Gilliland, K. O., & Kernick, E. T. (2014). Using Rasch Measurement to score, evaluate, and improve examinations in an anatomy course. *Anatomical Sciences Education*, *7*(6), 450-460. http://dx.doi.org/10.1002/ase.1436