Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 5 (2019) 82–89

www.elsevier.com/locate/hpe

A Simple Methodology for Discerning Item Construction Flaws in Health Professions Examinations

Kenneth D. Royal[a,*], Mari-Wells Hedgpeth[b], Lysa P. Posner[c]

[a]*Department of Clinical Sciences, North Carolina State University, 1060 William Moore Dr., Raleigh, NC 27607, USA*
[b]*Office of Academic Affairs, North Carolina State University, USA*
[c]*Department of Molecular Biomedical Sciences, North Carolina State University, USA*

Received 14 November 2017; received in revised form 5 February 2018; accepted 1 March 2018
Available online 7 March 2018

Abstract

Purpose: To introduce to health professions educators a simple methodology that can help discern item construction flaws and mitigate testwiseness effects.
Method: The methodology involved administering a veterinary medical school exam to medical school professional staff participants with no formal training in the medical and health sciences.
Results: The methodology was evidenced to be robust, as multiple items containing item construction flaws were identified by inflated success rates for a group of examinees who had no prior training in the subject matter.
Discussion: Health professions educators are encouraged to utilize the methodology presented in this paper, where appropriate, to discern item construction flaws, reduce measurement error, and increase score validity relating to their assessments.

Keywords: Medical education; Assessment; Psychometrics; Faculty Development; Grading; Health professions education

1. Introduction

In most medical and health professions programs, multiple-choice questions (MCQs) remain the most common method for assessing student performance. This largely is due to the efficiencies associated with the administration and scoring of exams provided to large class sizes, greater objectivity in grading, and increased score defensibility. Despite the advantages of MCQs, constructing high-quality items remains a persistent challenge for educators.[1,2] In fact, a recent study investigating the prevalence of item construction flaws at a large medical school in the United States found approximately 1 in 5 items contained a construction flaw.[3] The consequences of item flaws are quite severe as it increases measurement error that may threaten the validity of a score.[4] Given the moderate-to-high stakes associated with medical school assessments, it is critical that item flaws are minimized.[5]

Items with construction flaws often possess vulnerabilities that make them subject to being answered correctly by examinees with little, if any, subject matter knowledge. That is, examinees may be able to identify

*Corresponding author.
E-mail address: kdroyal2@ncsu.edu (K.D. Royal).

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2018.03.001

a correct answer by relying on good test-taking skills, a term known in the psychometrics literature as ‘testwiseness’.[6,7] Further complicating matters is the fact that most students have received some education about effective test-taking skills by the time they reach college, graduate school or other professional program (e.g., often as part of K-12 education, college study skills courses, and/or examination preparation efforts, etc.). At the heart of testwiseness is the ability to improve one's odds of successfully answering an item based on some guessing strategy. Rogers[8] identified three types of guessing strategies that examinees may utilize as a testwiseness strategy: random, cued and informed. Random guessing occurs when an examinee selects an item for no particular reason. Cued guessing occurs when an examinee detects a stimulus in the item that helps improve his or her chance of selecting the correct answer. Informed guessing (also known as an “educated guess”) occurs when an examinee has some partial knowledge of the subject and selects a response accordingly. In theory, random guessing would yield the least likelihood for success, whereas cued and informed guesses would lead to improved odds.

Clearly, students occasionally will rely on guessing strategies when attempting a multiple-choice item. While all examinees will have some chance of answering any given item correctly (e.g., an examinee attempting an item with four response options will have a 1 in 4 chance of success, at minimum), it is important that the likelihood of success is reduced to the smallest degree possible (e.g., one's chance of success based on random guessing) for those who do not know the correct answer. The problem, however, is how can an educator know if his/her items contain vulnerabilities that increase students’ odds of success beyond that of random guessing? To answer this question, the authors of this paper have devised a novel, yet simple methodology for discerning item construction flaws that can help mitigating testwiseness effects. Thus, the purpose of this paper is to introduce the methodology to health professions educators, describe how it works and demonstrate its utility for identifying item construction flaws that savvy test-takers might exploit to increase their examination performance.

2. Methods

2.1. Conceptual framework

An educated person, particularly someone with a minimum of a bachelor's (4-year college) degree, with no formal training in the medical and health sciences should possess a baseline level of knowledge (novice) about a specialized subject matter (e.g., medical content). Thus, such an individual would be expected to provide a series of random guessing strategies to attempt to pass a medical examination. A random guessing approach should result in a success rate of approximately 25% for items with 4 response options. However, an educated person would likely possess some reasonable amount of testwiseness skills due to at least 16 years of formal education and having participated in hundreds, or possibly thousands, of assessments over his or her educational career. An educated person would have a greater likelihood of identifying an item construction flaw and subsequently exploiting the flaw in order to identify the correct answer. Therefore, an educated person would have an increased likelihood of performing at a level greater than chance would dictate should the items possess any construction flaws that make would make them vulnerable to testwiseness strategies. Hence, an educated person should have a minimal level of content knowledge in the medical subject area, but also have the propensity to perform at a level greater than random chance due to possessing some degree of testwiseness skills. So, it stands to reason that if one can identify items that a collective group of individuals with no training in the subject area can answer with greater than 25% success (as determined by a difficulty estimate of .25), those items identified with higher success rates may possess some characteristic that makes it vulnerable to testwiseness strategies. These items, once identified, would make good candidates for review, and possible revision before administering again.

2.2. Participants

A convenience sample of professional staff employed in a college of veterinary medicine was invited to participate in the study. Participants were informed the purpose of the study was to determine how well similarly highly educated people in fields other than veterinary medicine could perform on an examination intended to measure one's fund of veterinary physiology knowledge. Staff personnel were employed in various departments, such as Academic Affairs, Student Affairs, Administration, Finance, Alumni Relations, Foundations, Educational Support Services, etc. For inclusion in the study, participants must hold at least a bachelor's degree and have no formal educational training in the medical or health sciences beyond the courses required as general education components of an undergraduate degree program. The purpose of this criterion was to attain a sample of similarly educated individuals with minimal medical (human and veterinary) content knowledge. Participants were informed that completing the examination was purely voluntary and in no way would affect their job status. A potential indirect benefit of the study was participants could request the result of their exam and claim bragging rights should one perform well on a veterinary medical school exam without having actually attended a single class in veterinary medical school. Participants also were made aware that all results would be treated with strict confidentiality and presented in aggregate form only. A total of 16 individuals were invited to participate in the study and 15 individuals agreed to participate and put forth their best effort to pass the exam. Permission to conduct the study was granted by the institution's Institutional Review Board (IRB) (Protocol #12226).

2.3. Course and instrumentation

An examination from a first year Veterinary Physiology II course served as the instrument for this study. The examination focused on renal physiology and was administered to 100 veterinary students in the spring of 2016 and 2017, respectively. Two cohorts were selected to ensure the psychometric properties of the items were both comparable and statistically stable. The examination consisted of 40 items. The 2017 examination had a KR-20 reliability coefficient of .832, with difficulty estimates (the percentage of students that answered the item correctly) ranging from .64 to 1.00. The mean difficulty estimate was .91 (SD = .08). Point biserial correlation estimates (a statistic used to determine if the most able examinees answered a given item correctly and if the least able students answered a given item incorrectly) ranged from 0.01 to .50 with a mean value of .25 (SD = .15). The 2016 examination had a KR-20 reliability coefficient of .823, with difficulty estimates ranging from .68 to 1.00. The mean difficulty estimate was .92 (SD = .07). Point biserial correlation estimates ranged from .05 to .58 with a mean value of .26 (SD = .19). An independent samples *t*-test indicated item statistics resulting from the 2016 and 2017 examinations did not differ with respect to difficulty, $t(78) = .446$, $p = .657$, or discrimination, $t(78) = .136$, $p = .892$, with alpha set at .05.

2.4. Procedures

Because differences in the conditions under which examinees complete an exam may serve as a source of measurement error, we attempted to replicate the procedures for staff personnel to the extent possible. Unfortunately, however, every aspect could not be replicated given the nature of the study. While the exam format (paper-and-pencil) and time constraints were identical for students and staff, we could not replicate the moderate-to-high stakes associated with the exam for students.

2.5. Analysis

The Classical Test Theory (CTT) framework[9,10] was utilized in the present study, and all psychometric indicators were calculated using Winsteps measurement software.[11] In particular, item difficulty, discrimination, and reliability coefficients were produced for both extant student data (years 2016 and 2017) and for the professional staff. Descriptive statistics of examinees' performance also were produced. Correlations were calculated to measure the association between difficulty values across the 2016, 2017 and staff personnel exams. Independent samples *t*-tests were performed to test for statistically significant differences between each set of exam statistics. SPSS statistical software (version 24.0) was used to perform all statistical analyses.

Baseline guessing error (BGE) refers to the difference between a group's actual performance minus expected performance.[12] For the present study, professional staff examinees were expected to achieve a collective success rate of 25% on each item (given 4 response options) as a result of utilizing random guessing strategies. The greater examinees' scores deviate from 25%, the larger the BGE. Similarly, the closer examinees' scores approximate their expected performance of 25%, the smaller BGE.

Finally, a ‘zone’ framework originally described in Royal and Hedgpeth[12] was used to determine the degree to which each item may possess a construction flaw vulnerability. These zones were used to qualitatively distinguish the degree to which each item posed a validity threat based on a presumed construction flaw. According to the evaluative framework, item difficulty estimates (p-values) between 25–33% would comprise the ‘low caution zone’, values between 34–50% would comprise the ‘caution zone’, and values exceeding 50% would comprise the ‘danger zone’. After data were analyzed, the results were presented to the faculty member responsible for the course for a qualitative analysis. The faculty member and the assessment team then worked together to theorize explanations for any questionable items.

3. Results

With respect to staff performance, overall percent correct scores ranged from 15–40%, with a mean score of 28.67% (SD = 6.40%) and a median of 30%. All items appearing on the exam contained four response options, which correspond to 25% odds of success for each item. The overall baseline guessing error (BGE) for the exam was 3.67%. BGE for individual items, however, were quite variable. Utilizing the 'zone' framework,[12] results indicate 37.5% of the exam contained items (n = 15) in the 'safe zone', 37.5% (n = 15) in the 'low caution zone', 12.5% (n = 5) in the 'caution zone' and 12.5% (n = 5) in the 'danger zone'.

Given the exam was not properly targeted to staff personnel ability levels, it was expected the psychometric properties would appear vastly different from those produced by students interaction with the exam. The KR-20 reliability coefficient was .00, and item difficulty values ranged from .00 to 0.67 with a mean of .29 (SD = .15). An independent samples *t*-test revealed the differences in difficulty between the 2016 and staff exams were statistically significantly different, $t(78) = 23.550$, $p = .000$, as were the differences in difficulty between the 2017 and staff exam, $t(78) = 23.031$, $p = .000$. Discrimination coefficients were not compared due to the extremely poor person/item targeting that would threaten the validity of any resulting values. The correlations between the difficulty values obtained from 2016 and 2017 student exams were $r = .936$ ($p = .000$) and $\rho = .892$ ($p = .000$). The correlations between the 2016 student exam and the professional staff exam were $r = .221$ ($p = .170$), $\rho = .209$ ($p = .196$). The correlations between the 2017 student exams and the professional staff exam were $r = .169$ ($p = .296$), $\rho = .211$ ($p = .191$).

4. Discussion

4.1. Substantive findings

Given staff participants had no formal educational training on veterinary renal physiology, it is almost assured that participants utilized some primary combination of random and cued guessing strategies to complete the exam. Effective use of cued guessing strategies can eliminate distractors, thus improving one's odds of selecting the correct answer. In theory, participants with no training in veterinary renal physiology would have a 25% chance of answering each item correctly and culminating in an overall score of 25%. In the present study, participants collectively answered 28.67% correct on average, with 3 participants scoring less than 25%, 2 participants scoring 25%, and 10 participants scoring slightly better than 25%. The lowest score was 15% and the highest score was 40%. Further, item difficulty correlations indicated a strong association between the student exams administered in 2016 and 2017, and a very weak association between each student exam and the professional staff exam. Thus, it was clear the exam functioned very differently for staff personnel.

With respect to understanding the degree to which the items were vulnerable to testwiseness strategies, both person and item results are assuring. However, the collective group answered 10 items (25% of the exam) with greater than 34% success. These items are prime candidates for potential revision before re-administering to students, as the higher success rates imply staff participants were able to improve their odds of success from 1 in 4 to at least 1 in 3. Typically, greater than expected performance is a signal of a poorly functioning distractor(s) that may need revision. Educators are encouraged to review item distractor statistics for the staff exam first to diagnose the problem. Because a false-positive detection may occur if the distractor does not show similarly poor functioning on students' exams, educators should compare item distractor statistics to discern if the flagged items possess true vulnerabilities, or simply are an artifact of a series of collective 'lucky guesses'.

Another approach to help discern if flagged items are problematic is to graph the item difficulty statistics (see Fig. 1). Student exams should present similar trend lines, but the staff exam should contrast sharply. Instances in which easier items for students are also relatively easy for staff likely is an indicator of very poor distractors, and calls into the question the validity of the responses obtained for these items (e.g., 1, 21, and 27). Likewise, easy items for students that are difficult for staff (e.g., 2, 11, and 26) might present some validity evidence to support the fidelity of those item responses. Conversely, difficult items for students that also are difficult for staff (e.g., 25) might present some validity evidence to support the fidelity of those item responses, whereas difficult items for students that were relatively easy for staff (e.g., 9) may likely possess either excellent distractors for a discerning student or tricky distractors that confuse potentially knowledgeable students.

Findings from this study also are assuring with respect to student learning, as it is clear that some high level of medical content knowledge is necessary to pass the exam. However, it should be noted that most

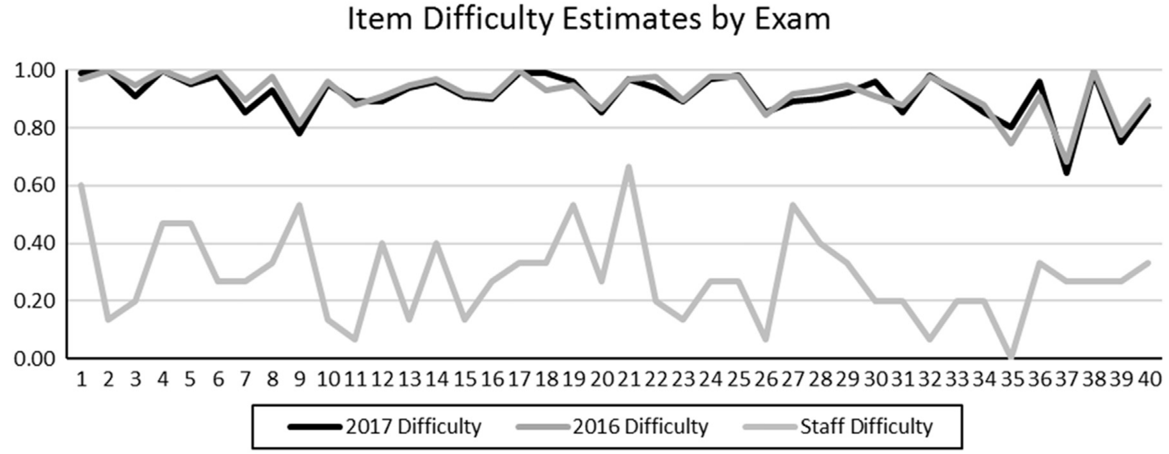

Fig. 1. Comparison of item difficulty estimates by exam.

examinees occasionally rely on some guessing strategy (typically informed guessing) to select responses. Thus, most exam scores will possess some inescapable element of measurement error. Relatedly, one might also wish to know whether or not students' exam performance likely is evidence of authentic learning. To answer this question many factors should be considered. For example, educators should consider the origins of the exam items and the manner in which instruction was provided. These reflections can help educators understand why students performed so well on some items, and perhaps not so well on others. Insights gleaned from this reflection can help inform future teaching opportunities.

After results were produced and the findings were shared with faculty member responsible for the course, we collaborated to identify reasons why some items may contain some vulnerabilities to testwiseness. Below, we present three items with elevated success rates for staff personnel and provide our reasoning for discerning the type of flaw.

Example #1. : Where is aldosterone produced and stored? (#21 in Fig. 1)

A. Kidney
B. Adrenal gland
C. Liver
D. Lung

For this item, 10 of 15 (66.7%) staff participants selected B, which is the correct answer (3 selected A, and only one selected C and D, respectively). It is likely that participants were able to identify the correct answer due to possessing some partial knowledge of the subject, thus utilizing an informed guessing strategy.

Example #2. : Approximately, what percentage of cardiac output feeds the kidneys? (#1 in Fig. 1)

A. 10%
B. 25%
C. 50%
D. 80%

For this item, 8 out of 15 (53.3%) participants selected B, which is the correct answer (1 person selected A, 3 selected C and 1 selected D). It is likely that participants would have to rely on a random guessing strategy given the item is unlikely to provoke partial knowledge among participant, nor is the item flawed so it is unlikely to provoke a cued guessing strategy. In all likelihood, this item was answered correctly due to "edge aversion", which is the notion that examinees will tend to avoid the extremes when presented a range.[13] To counteract this bias towards this middle an educator might make the distractors more similar to the correct answer (e.g., 20%, 30%, 35%).

Example #3. : Under normal everyday conditions, the GFR is kept steady in face of alterations in blood pressure by which of the following mechanisms? (#9 in Fig. 1)

A. Actions of ADH
B. Actions of aldosterone
C. The osmoregulatory system
D. The responsiveness of the afferent and efferent arterioles to local changes in blood flow/pressure

For this item, 8 out of 15 (53.3%) participants selected D, which is the correct answer (1 person selected A, 3 selected B and C, respectively). In this example, the item's author made the correct answer the longest. This is likely due to the author writing the question and correct answer first, and then trying to generate three other plausible choices. Testwise examinees typically can detect this cue and use a cued guessing strategy to identify the correct answer.

4.2. Implications

With respect to implications, perhaps the most obvious is that the methodology may serve as an additional source of validity evidence,[14,15] as staff personnel results should diverge from students' scores. Another interesting byproduct of this study is that it may produce a baseline estimate of how well an equivalently educated person with no formal training in veterinary physiology can be expected to perform in a given content area. This baseline estimate can be of value when attempting to distinguish learning gains acquired from a course.

4.3. Other considerations

A psychometric review of examination functioning and student performance is highly advisable in medical and health professions educational assessments given the moderate-to-high stake nature of most exams. Such analyses should occur immediately after every major exam and prompt efforts should be made to clarify anything that remains unclear to students when possible. It should be noted, however, that clarification efforts should involve a substantive review of content, and not a re-exposure to actual items.[16] Of course, items with discernible flaws should always be corrected before re-administration, as flawed items contributes measurement error that threatens score validity. It also is advisable to consider pilot testing new items as unscored, experimental items. This allows educators the opportunity to ensure a new item is psychometrically sound and functioning properly before it carries any stakes for students. It also provides educators with insights about the difficulty of the item, which can be particularly useful when constructing/revising an exam to maintain a comparable level of difficulty.

The exercise presented in this paper describes a simple, common-sense approach that can help educators understand the degree to which their exam items potentially possess psychometric flaws. Practically speaking, however, it is unadvisable to subject most staff personnel to too many exercises such as this one, as they may grow frustrated or disappointed by their performance. Therefore, educators should utilize exercises such as the one described in this paper judiciously if involving teams that have limited involvement with assessment and curriculum support activities. We opted to test this validation approach on an exam that was selected out of convenience. In the event we conduct this exercise again with staff from numerous departments we likely will 1) draw a sample of items from multiple exams (or courses) to cover the broadest scope possible; or 2) identify only the items that were deemed easiest for students (e.g., difficulty estimates greater than 90%) in order to test if the high success rates may indeed be attributed to excellent teaching and learning, good guessing, or some combination of the two.

Of course, for those staff personnel that work most closely with faculty educators (e.g., assessment staff, educational technology staff, curriculum support staff, etc.), many (if not all) may will be willing to complete each major exam (e.g., mid-terms, finals, etc.) as part of an effort to mitigate testwiseness advantages. Because exams would not need to be completed by staff personnel each semester/year, a reasonable implementation strategy might involve completing major assessment once every 3–5 years, or perhaps evaluating only new and/or recently revised items. Over time, such a process would ensure that virtually every item intended for use has been vetted as part of the external validity check. The insights obtained from such analyses go well beyond item construction reviews and may offer insights undiscernible via traditional flaw analysis exercises.

Some might consider the methodology presented in this paper as somewhat controversial as it involves utilizing staff personnel for a task that typically would fall beyond the scope of one's normal duties. It is possible that some staff personnel may experience feelings of coercion if asked by a supervisor to participate in a project of this nature. Thus, it is important to work with one's institutional review board (IRB) to ensure staff participants' rights are protected. Part of the ethics associated with this type of exercise requires informing staff that their participation is purely voluntary and not an expectation of employment. Staff should also have the option to withdraw from the study at any time. It has been our experience, however, that staff generally are very intrigued by the idea and are eager to assist. In fact, only 1 of 16 individuals invited to participate in the study declined. Many indicated they have often wondered how well they could perform on a medical school exam and viewed the exercise as a fun and safe opportunity to find out. For many staff, this exercise also offers a whole new appreciation for the sophisticated content that veterinary students must learn. Given staff likely will perform poorly, it is a good idea to inform them that poor performance is expected because they have not attended veterinary medical school or otherwise received any training in the content area. This will help set appropriate expectations for participants and help avoid feelings of disappointment if/when one's score performance is low. Of course, one should also treat all data with strict confidentiality and not reveal any information that could link an individual to his/her scores. While a blinded summary of score results may be appropriate to show participants how the collective group performed, never should one reveal which individual received each score as it could potentially embarrass participants and violate trust.

4.4. Limitations

There are several limitations of this study. Perhaps the most obvious is the small sample size. The potential statistical instability resulting from only 15 participants could impact the degree to which items accurately fit into 'zones'. It is possible that a larger sample could cause some items to shift from more to less concerning zones, and vice versa. Another potential limitation involves participants' motivation levels, as staff participants clearly did not experience the same levels of pressure as students to perform well. Additionally, we were unable to utilize more robust psychometric procedures (e.g., item response theory, Rasch models, etc.) due both to the small sample size and the high likelihood of significant misfit given the exam was not appropriately targeted for a group of examinees with no training in the subject matter.[17] Finally, we are unaware of the specific strategy participants used to answer each item. Having insights about whether participants used random, cued and informed guessing strategies could help further discern item vulnerabilities.

5. Conclusions

The purpose of this paper was to introduce to health professions educators a simple methodology that can help discern item construction flaws and mitigate testwiseness effects. The methodology involved testing 15 staff participants with an examination that was administered to veterinary medical students for the past two years. The methodology proved robust, as multiple items containing item construction flaws were identified that resulted in inflated success rates for a group of test-takers who had no prior training in the subject matter. Health professions educators are encouraged to utilize the methodology presented in this paper, where appropriate, to discern item construction flaws, reduce measurement error, and increase score validity.

One sentence bios

Royal is Assistant Professor of Educational Assessment & Outcomes and Co-Director of the Office of Assessment, Evaluation and Research at the North Carolina State University College of Veterinary Medicine.

Hedgpeth is Co-Director of the Office of Assessment, Evaluation and Research at the North Carolina State University College of Veterinary Medicine.

Posner is Professor of Anesthesiology, Assistant Department Head, and Director of Anesthesia Services at the North Carolina State University College of Veterinary Medicine.

Financial support

None.

Conflicts disclosure

The authors declare no conflicts of interest

Ethical approval

Ethical approval was granted by the institution's Institutional Review Board (IRB), Protocol #12226

References

1. Dragsow F, Luecht RM, Bennett R. Technology and testing. In: Brennan RL, editor. *Educational Measurement*, 4th edition, Washington, DC: American Council on Education; 2006. p. 471–516.
2. Rodriguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educ Meas* 2005;24(2):3–13.
3. Royal KD, Hedgpeth WM. The prevalence of item construction flaws in medical school examinations and innovative recommendations for improvement. *Eur Med J: Innov* 2017;1(1):61–66.
4. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ* 2005;10(2):133–143.
5. American Educational Research Association. *American Psychological Association, & National Council on Measurement in*

Education. Washington, DC: Standards for Educational and Psychological Testing; 2014.

6. Thorndike RL. Reliability. In: Lindquist EF, editor. *Educational Measurement*. Washington DC: ACE; 1951. p. 560–620.
7. Millman J, Bishop H, Ebel R. An analysis of test-wiseness. *Educ Psychol Meas* 1965;25:707–726.
8. Rogers HJ. Guessing in multiple-choice tests. In: Masters GN, Keeves JP, editors. *Advances in Measurement in Educational Research and Assessment*. Oxford, UK: Pergamon; 1999. p. 235–243.
9. Allen MJ, Yen WM. *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press; 2002.
10. Novick MR. The axioms and principal results of classical test theory. *J Math Psychol* 1966;3(1):1–18.
11. Linacre JM. WINSTEPS® (Version 3.92.0). Computer Software. Beaverton, OR: Winsteps.com; 2017.
12. Royal KD, Hedgpeth MW. A novel method for evaluating examination item quality. *Int J Psychol Stud* 2015;7(1):17–22.
13. Attali Y, Bar-Hillel MB. Guess where: the position of correct answers in multiple-choice test items as a psychometric variable. *J Educ Meas* 2003;40(2):109–128.
14. Kane MT. Validation. In: Brennan R, editor. *Educational Measurement*, 4th ed., Westport, CT: Praeger; 2006. p. 17–64.
15. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. *Adv Med Educ Pract* 2017;8:567–570.
16. Royal KD, Henderson A, Hedgpeth WM. Post-exam reviews: a consideration of costs and unintended consequences. *Med Sci Educ* 2015;25:327–329.
17. Royal KD, Gilliland KO, Kernick ET. Using Rasch measurement to score, evaluate, and improve examinations in an anatomy course. *Anat Sci Educ* 2014;7(6):450–460.