Original Article

# A Method for Investigating "Instructional Familiarity" and Discerning Authentic Learning

**Royal KD, Hedgpeth MW[1], Smith KW[2], Kirk D[3]**

*Department of Clinical Sciences, North Carolina State University, [1]Office of Veterinary Medical Education, North Carolina State University, Raleigh, [2]Office of Medical Education, University of North Carolina, [3]Department of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA*

**Address for correspondence:**
Dr. Kenneth D Royal,
1060 William Moore Drive,
CVM Main Building, C-296, Raleigh,
NC 27607, USA.
E-mail: kdroyal2@ncsu.edu

## Abstract

**Background:** Presently, most medical educators rely exclusively on item difficulty and discrimination indices to investigate an item's psychometric quality and functioning. We argue "instructional familiarity" effects should also be of primary concern for persons attempting to discern the quality and meaning of a set of test scores. **Aim:** There were four primary objectives of this study: (1) Revisit Haladyna and Roid's conceptualization of "instructional sensitivity" within the context of criterion-referenced assessments, (2) provide an overview of "instructional familiarity" and its importance, (3) reframe the concept for a modern audience concerned with medical school assessments, and (4) conduct an empirical evaluation of a medical school examination in which we attempt to investigate the instructional effects on person and item measures. **Subjects and Methods:** This study involved a medical school course instructor providing ratings of instructional familiarity (IF) for each mid-term examination item, and a series of psychometric analyses to investigate the effects of IF on students' scores and item statistics. The methodology used in this study is based primarily on a mixed-method, "action research" design for a medical school course focusing on endocrinology. Rasch measurement model; correlation analysis. **Results:** The methodology presented in this article was evidenced to better discern authentic learning than traditional approaches that ignore valuable contextual information about students' familiarity with exam items. **Conclusions:** The authors encourage other medical educators to adopt this straightforward methodology so as to increase the likelihood of making valid inferences about learning.

**Keywords:** Action research, Assessment, Medical education, Psychometrics, Testing

## Introduction

In 1981, Haladyna and Roid[1] published an important paper in the Journal of Educational Measurement on the topic of "instructional sensitivity." The authors defined instructional sensitivity as "the tendency for an item to vary in difficulty as a function of instruction" (p. 40). They go on to say "When instruction is reasonably effective, items given to uninstructed students should appear difficult and the very same items when administered to instructed students should appear easy" (p. 40). Although, Haladyna and Roid were not the first researchers to address the topic of instructional sensitivity,[2-7] they were among the first to argue instructional sensitivity was a critical component of criterion-referenced assessments and should be an assessor's primary concern when attempting to assess instructional effects on student learning.

Fast-forward 30 years and the topic of instructional sensitivity has once again gained considerable attention in the mainstream educational research literature. However, the present definition has evolved, and the topic is primarily discussed

| Access this article online | |
|---|---|
| Quick Response Code: | Website: www.amhsr.org |
| | DOI: ***** |



**How to cite this article:** ???

in the context of large-scale K-12 assessments. Presently, most accept the definition provided by renowned educational researcher Popham that states "A test's instructional sensitivity represents the degree to which students' performances on that test accurately reflect the quality of instruction specifically provided to promote students' mastery of whatever is assessed" (p. 146).[8] The purpose of this article is to (1) revisit Haladyna and Roid's conceptualization of instructional sensitivity within the context of criterion-referenced assessments, (2) provide an overview of its importance, (3) reframe the concept for a modern audience concerned with medical school classroom assessments, and (4) conduct an empirical evaluation of a medical school examination in which we attempt to investigate the instructional effects on person and item measures.

### Instructional familiarity

As noted previously, the definition of instructional sensitivity has evolved over the past several decades. With the rising popularity of criterion-referenced examinations, instructional sensitivity was generally assumed, as opposed to tested.[9] Those that did test instructional sensitivity throughout the 1980s and the 1990s generally referred to the term as "instructional validity,"[10] as the term was meant to distinguish instructional effects as a property of instructional quality. Throughout the 1970s and 1980s, the term "curricular sensitivity"[11,12] was also used to indicate the effects of instruction upon a curriculum. This term was intended to distinguish instructional effects as a property of curricular quality. In any instance, it is clear that the term has experienced a great deal of turbulence over the last several decades. Therefore, we propose to modernize Haladyna and Roid's conceptualization as instructional familiarity (IF), as we contend it is a student's familiarity with instructional content (and items) that can cause items to vary in difficulty. Under this conceptualization, the influence of IF directly impacts an examination score, thus making IF a testable psychometric property.

We believe IF is of critical importance to persons concerned with the discernment of authentic learning and making valid score inferences. Presently, most medical educators rely exclusively on item difficulty and discrimination indices to investigate an item's psychometric quality and functioning. We agree with Haladyna and Roid's thesis that IF effects should be a primary concern when attempting to evaluate a set of examination scores and understand why various items function as they do. We also contend that only by understanding the influence of IF on a set of examination scores can one truly begin to understand the extent to which scores are authentic evidence of student learning, or merely an artifact of scores resulting from "teaching to the test."

## Subjects and Methods

### Study design

We attempted to investigate the potential effects that instructionally familiar material may have a set of moderate-to-high stakes medical school examination scores. The educational research-based "action research" methodology involved having the course instructor provide a rating of IF for each item before the administration of the examination. Upon administration of the examination, data were analyzed by both classical test theory and Rasch measurement frameworks to better understand the many nuances of the data. Next, the relationship between the instructor's IF ratings and students' performance on each item was investigated to determine the extent to which potential IF effects could have some bearing on students' examination scores. With regard to human subjects' projection and research ethics, this study was deemed exempt.

### Course background

Endocrine system and nutrition are a 2½ weeks course in the 2nd year of the MD Program at The University of North Carolina School of Medicine. Content pertaining to the normal development, structure, and function of the endocrine system as well as the pathophysiology, clinical manifestations, diagnosis, and basic management of hormone excess/deficiency states is covered during the course. The didactic course material is presented mainly through large group lecture sessions and seven small group discussions based on clinical cases. The course is divided into two thematic sections. The first section covers diseases of the major hormone-producing organs: The hypothalamus, pituitary, adrenal, thyroid, and parathyroids including calcium and bone disorders. A common thread is introduced for each organ in this section, starting with a review of normal function and histology, then moving to the clinical presentation and management of the associated diseases, and finally solidifying understanding of lab interpretation. The second section of the course is devoted to glucose homeostasis, diabetes, energy balance, and obesity.

### Sample

The sample frame for this course consisted of 182 second year medical students. The gender was represented with 54% (98/182) males and 46% (84/182) females. Students ranged in age from 22 to 37, with a mean (standard deviation [SD]) age of 26 (2.6) and a median age of 25. With respect to race, 63% (114/182) were White/Caucasian, 13% (24/182) were African American, 7% (13/182) were Asian Indian, 7% (13/182) were Chinese, and others including Japanese, Korean, Native American, and Filipino, and 10% (18/182) did not indicate their race.

### Examination

A multiple-choice mid-term examination served as the apparatus for this study. The exam contained 43 items that were intended to measure students' medical and clinical knowledge of the normal and abnormal pituitary, adrenal, thyroid, and parathyroid glands that were covered in the first 6 days of the course. All items were in the "single-best answer" format with 1 item (2%) (1/43) having three response options, 36 of 43 (84%) having four response options, and 6 of 43 (14%) having five response options. All items were clinical in nature

and contained short vignettes, lab results, and/or graphics that focused on the application, evaluation, and synthesis of knowledge. The examination was administered during an hour and a half period in a series of secure examination rooms with 36 students/room. Students completed the examination using laptop computers and the web-based Medical Student Testing and Report System (MedSTARS) application. MedSTARS is an online application developed at the University of North Carolina Medical School and features a secure login and verification. For this particular examination, it was used in conjunction with an online proctoring application that allows student laptops to be locked down (to prevent the use of other online and internet resources) and monitored by a human proctor.

### Instructional familiarity scale

For the purpose of our study, we created an instrument named the IF Scale (IFS). This instrument was used by the course instructor to determine the extent to which instructionally familiar material was presented on each mid-term examination item. The IFS provided a five-point continuum onto which the course instructor could rate each item from not instructionally familiar (1) to very instructionally familiar (5). Some factors that may contribute to IF include: (1) The extent to which the item, or its content, may be recognizable to examinees; (2) the amount of time spent in class, small groups, and outside required work on a particular content domain; (3) the extent to which the instructor emphasized content during the course; (4) the source of the item (e.g., required textbook/reading/assignment or supplementary material such as a suggested outside reading list); and (5) the extent to which the instructor implied particular content would appear on the exam. It is important to note these factors are not exhaustive, but do illustrate some common reasons why an item may or may not be familiar to examinees.

### Procedures

In this study, the instructor was provided the aforementioned operational definition of IF and presented with examples of various sources from which IF might emanate. The instructor was asked to thoughtfully reflect on this concept and use her best judgment to determine the extent to which various examination items might be instructionally familiar to students. In addition to the directions (which included the conceptualization and mental calibration exercise), the instructor was provided an Excel Spreadsheet that was prepopulated with a column containing the item numbers for each of the 43 item (presented in the order as they would ultimately appear on the exam) and a column indicating where to record the IF rating (1–5). Upon administration of the mid-term examination, instructor ratings were merged into a common data set with various testing data for analysis.

### Data analysis

Data analysis involved both a classical test theory and Rasch analysis of data. The psychometric properties of the mid-term examination were investigated to ensure the examination possessed adequate psychometric properties. Both item statistics and IF rating distributions were evaluated, as well the relationship between IF ratings and item difficulty estimates. Rasch-based person/item maps provided an additional visual of the results. SPSS (IBM Corp. Armonk, NY) and Winsteps measurement software[13] were used to perform the various analyses.

## Results

### Psychometric indicators of quality

The mid-term examination was subjected to a Rasch analysis for quality control and data analyses. Results indicate the examination had a reliability of 0.73, and a separation estimate of 1.65 indicating about 2.53 strata or statistically distinguishable levels of item difficulty.[14] Overall, the data fit the Rasch model[15] very well with an infit mean square estimate of 1.00 and an outfit mean square estimate of 0.99 for both persons and items. Rasch analyses use two primary indicators to inform item quality: Discrimination (point-measure correlations) and fit statistics (infit and outfit mean square statistics). All items indicated positive point-measure correlations and all items demonstrated infit and outfit mean square fit statistics between the recommended range of 0.7 and 1.3 for an examination with these stakes.[16] Collectively, results indicate the examination was psychometrically sound and functioned appropriately.

### Examination results

A total of 182 students completed each of the 43 mid-term examination items. The mean (SD) student score was 77% (12%) with a range of 47–100%. The mean (SD) item *P* value (percent correct) was 0.77 (0.12), and ranged from 0.42 to 0.94. The minimum passing standard for the examination was 60%, which resulted in 13 students failing the examination.

### Instructional familiarity results

The course instructor provided an IF rating for each of the 43 items. The Spearman's rho correlation between IF ratings and item *p*-values was 0.28, indicating a negligible to the weak relationship. However, it is important to note that some of the data contained outliers and extreme scores that could influence the correlation estimates. Such examples included item number 39, which had an IF rating of 5, meaning the content should have been very instructionally familiar to students, but only 56% of students answered the item correctly. Other instances of extreme scores were found in the IF rating category of 2, where 3 items had unexpectedly high *p*-values ranging from 0.89 to 0.92.

To better understand the relationship between IF ratings and item difficulty, a series of analyses were performed. First, a breakdown of descriptive statistics for each categorical rating is presented in Table 1.

Next, we present a full breakdown of item statistical performance by IF rating in Table 2. Results are sorted according to IF rating, with the least familiar items at the top and the most familiar items at the bottom. Item *p*-values indicate the percentage of students that answered each item correctly. It appears, there are three discernible levels of student performance, with IF ratings of 1, 2, and collapsed ratings of 3–5.

Another method to investigate results involved rank-ordering items by their degree of difficulty and dividing the sample in half to see how rating fared between the two groups. Because Rasch analyses begin the item estimation process by centering items around the mean, the difficulty calibrations produced from the analysis provided the perfect basis for drawing the line between more difficult (any item with a positive logit value) and less difficult items (any item with a negative logit value). A total of 21 items comprised the easier set of items, and the remaining 22 items comprised the more difficult set of items. The results are presented in Table 3.

Finally, we investigated the results of the Rasch-based person/item maps (also called "Wright maps") as an additional (and visual) lens for interpreting results.

The person and item map presented in Figure 1 illustrates the psychometric ruler onto which measures of student performance and item difficulty are mapped so as to better understand how each functions separately, as well as relative to the other. For a detailed overview for interpreting the map, readers are referred to Royal.[17] In short, the left half of the map illustrates measures of student knowledge, and the right half of the map illustrates measures of each item's difficulty. The top of the chart indicates more of the latent trait (in this case, content knowledge of the endocrine system) for persons and more difficult items. Conversely, the bottom of the map indicates less of the latent trait is evidenced among persons, and less difficult items. Thus, item 35 was the most difficult item appearing on the examination, and item 28 was the easiest. One may specifically investigate the location of each item on the map according to its IF rating, or potentially color-code items by rating. In general, more instructionally familiar items tend to fall at the bottom of the ruler, and less instructionally familiar items tend to fall at the top.

Not all items performed as predicted by their IF ratings. In fact, there were 6 items (about 14% [6/43] of the total item pool) that performed in surprising ways [Table 4]. In particular, items 39, 30, and 23 were demonstrated to be more difficult than anticipated given the amount of familiarity students should have had with the content. Items 27, 21, and 17 were easier than anticipated given the familiarity students should have had with the content. To better understand why these items performed in unexpected ways, the instructor was asked to provide a qualitative review of each item outlier.

**Table 1: Descriptive statistics for each rating category**

| IF rating | Count | Mean *P* | SD | Median *P* |
|---|---|---|---|---|
| 1 | 3 | 0.72 | 0.05 | 0.70 |
| 2 | 13 | 0.73 | 0.15 | 0.74 |
| 3 | 12 | 0.82 | 0.07 | 0.82 |
| 4 | 11 | 0.80 | 0.12 | 0.82 |
| 5 | 4 | 0.77 | 0.16 | 0.80 |

SD: Standard deviation, IF: Instructional familiarity

**Table 2: Item statistics**

| Item number | Rasch difficulty | SE | Point-measure correlation | Raw score | *p* | IF rating |
|---|---|---|---|---|---|---|
| 36 | 0.67 | 0.17 | 0.21 | 124 | 0.68 | 1 |
| 31 | 0.58 | 0.17 | 0.42 | 127 | 0.70 | 1 |
| 33 | 0.10 | 0.19 | 0.30 | 142 | 0.78 | 1 |
| 35 | 1.93 | 0.16 | 0.34 | 76 | 0.42 | 2 |
| 10 | 1.39 | 0.16 | 0.41 | 97 | 0.53 | 2 |
| 6 | 1.18 | 0.16 | 0.37 | 105 | 0.58 | 2 |
| 2 | 1.07 | 0.16 | 0.43 | 109 | 0.60 | 2 |
| 15 | 0.52 | 0.17 | 0.42 | 129 | 0.71 | 2 |
| 19 | 0.33 | 0.18 | 0.22 | 135 | 0.74 | 2 |
| 43 | 0.33 | 0.18 | 0.38 | 135 | 0.74 | 2 |
| 41 | 0.20 | 0.18 | 0.32 | 139 | 0.76 | 2 |
| 11 | 0.06 | 0.19 | 0.33 | 143 | 0.79 | 2 |
| 26 | −0.47 | 0.22 | 0.13 | 156 | 0.86 | 2 |
| 17 | −0.79 | 0.24 | 0.29 | 162 | 0.89 | 2 |
| 21 | −0.92 | 0.25 | 0.29 | 164 | 0.90 | 2 |
| 27 | −1.21 | 0.28 | 0.13 | 168 | 0.92 | 2 |
| 13 | 0.49 | 0.17 | 0.21 | 130 | 0.71 | 3 |
| 29 | 0.49 | 0.17 | 0.34 | 130 | 0.71 | 3 |
| 40 | 0.24 | 0.18 | 0.36 | 138 | 0.76 | 3 |
| 3 | 0.17 | 0.18 | 0.12 | 140 | 0.77 | 3 |
| 16 | 0.03 | 0.19 | 0.19 | 144 | 0.79 | 3 |
| 7 | −0.09 | 0.20 | 0.30 | 147 | 0.81 | 3 |
| 42 | −0.16 | 0.20 | 0.16 | 149 | 0.82 | 3 |
| 9 | −0.33 | 0.21 | 0.30 | 153 | 0.84 | 3 |
| 1 | −0.52 | 0.22 | 0.18 | 157 | 0.86 | 3 |
| 24 | −0.85 | 0.25 | 0.21 | 163 | 0.90 | 3 |
| 34 | −0.98 | 0.26 | 0.23 | 165 | 0.91 | 3 |
| 32 | −1.29 | 0.29 | 0.18 | 169 | 0.93 | 3 |
| 23 | 1.15 | 0.16 | 0.45 | 106 | 0.58 | 4 |
| 30 | 1.07 | 0.16 | 0.31 | 109 | 0.60 | 4 |
| 38 | 0.27 | 0.18 | 0.38 | 137 | 0.75 | 4 |
| 5 | −0.09 | 0.20 | 0.32 | 147 | 0.81 | 4 |
| 20 | −0.09 | 0.20 | 0.18 | 147 | 0.81 | 4 |
| 4 | −0.16 | 0.20 | 0.21 | 149 | 0.82 | 4 |
| 37 | −0.38 | 0.21 | 0.28 | 154 | 0.85 | 4 |
| 8 | −0.52 | 0.22 | 0.28 | 157 | 0.86 | 4 |
| 12 | −0.92 | 0.25 | 0.22 | 164 | 0.90 | 4 |
| 14 | −0.92 | 0.25 | 0.14 | 164 | 0.90 | 4 |
| 18 | −1.38 | 0.30 | 0.24 | 170 | 0.93 | 4 |
| 39 | 1.26 | 0.16 | 0.27 | 102 | 0.56 | 5 |
| 22 | 0.06 | 0.19 | 0.21 | 143 | 0.79 | 5 |
| 25 | −0.05 | 0.19 | 0.27 | 146 | 0.80 | 5 |
| 28 | −1.48 | 0.32 | 0.22 | 171 | 0.94 | 5 |

SD: Standard error, IF: Instructional familiarity

**Table 3: Count of IF ratings relative to item difficulty**

| Examination difficulty | IF ratings | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| More difficult half (*n*=22) | 3 | 9 | 5 | 3 | 2 |
| Easier half (*n*=21) | 0 | 4 | 7 | 8 | 2 |

IF: Instructional familiarity

**Table 4: Qualitative review of items with IF and difficulty discrepancies**

| Item number | IF rating | *p* | Explanation |
|---|---|---|---|
| 39 | 5 | 0.56 | Question was presented differently than it was presented to students in class; adverse weather may also have limited student exposure |
| 30 | 4 | 0.60 | This item contained the same graph presented in lecture but asked a very specific question that was only addressed in small group |
| 23 | 4 | 0.58 | Concept presented in item is counterintuitive, and only the highest performing students tend to answer this item correctly |
| 27 | 2 | 0.92 | This item contains a factoid that is likely easily recalled, although mentioned only once in lecture |
| 21 | 2 | 0.90 | This item contains a factoid that is likely easily recalled, although mentioned only once in lecture |
| 17 | 2 | 0.89 | Specific content is not discussed, but the global content area is discussed at great length |

IF: Instructional familiarity

Collectively, there is some evidence that items generally performed in somewhat predictable ways based on their potential familiarity to students. However, the relationship between item performance and IF is, so modest that it only suggests some potential score contamination likely due to memory recall effects. Because the instructor did not "teach to the test," the scores resulting from this study are likely to have limited familiarity effects, and consequently, should result in stronger evidence of authentic learning.

## Discussion

Our results indicate there was a discernible difference in collective student performance on items with varying degrees of IF. In particular, there appeared to be three distinct levels of student performance: (1) IF ratings of 1; (2) IF ratings of 2; and (3) collapsed IF ratings of 3–5. There is some evidence suggesting the more familiar the content or item may be to students, the better they will perform. In the context of this study, this effect may be considered a positive or a negative. On the positive side, it is expected that students will perform better on items that are more familiar to them. However, on the negative side, it might also suggest that not all scores indicate authentic evidence of learning, as there may be some score contamination due to familiarity effects. Contamination

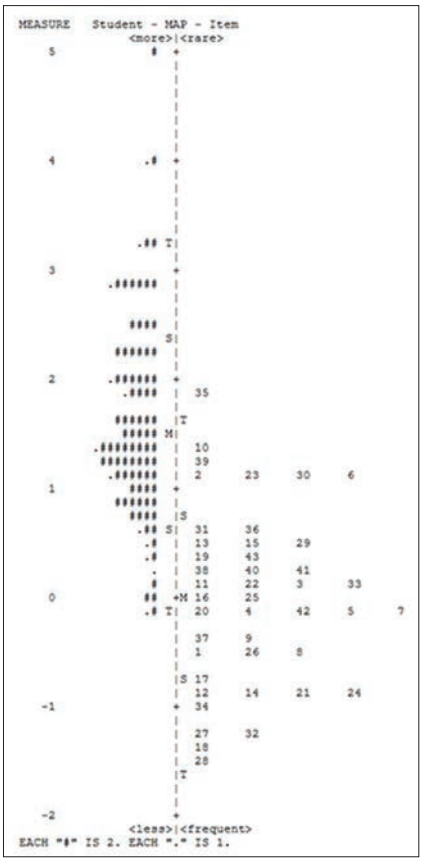


**Figure 1:** Person and item map

could result from memory and recall ability, psychological cues associated with the circumstances surrounding the delivery of the material, and so on, thus resulting in a potentially inaccurate reflection of what students truly know.

It is important to bear in mind that correlation values may not be the best measure for discerning IF. Correlations are extremely sensitive to outliers, and most examination data sets are somewhat "messy" as participants often respond in unexpected ways (e.g., guessing, correctly/incorrectly answering questions that yield a low/high probability of success relative to the person's ability, etc.). Further, a few ratings on the extreme ends of the IF rating scale provide very little power for accurately detecting the magnitude of a relationship. In the present study, the Spearman's rho correlation between IF ratings and item *P* values was 0.28, indicating a negligible to weak relationship. However, further inspection of the data by way of additional analyses revealed a much more informative perspective. We encourage readers to emulate many of the rudimentary methods presented in this study to better understand IF effects.

Despite the problems potentially associated with correlations, these estimates may certainly be of value when outliers are trimmed or removed from the data set. There remains the question of what a low, moderate, or high correlation may suggest. We contend that a high correlation might suggest very

little evidence of authentic learning and plenty of evidence that students can perform well on items that are familiar to them. Of course, such an inference can never be absolute or made without more information, but a high correlation would indicate a pattern that should be carefully examined. A moderate correlation might suggest some evidence of both authentic learning and score contamination due to familiarity effects. A low correlation might provide greater evidence of authentic learning, and minimal influence of IF effects. To be clear, even a very low correlation would not necessarily provide definitive evidence that students' scores are entirely uncontaminated by IF effects. It would, however, seem reasonable that a lower correlation is useful for discerning authentic learning.

Some outliers can be expected. In this study, 6 items performed unexpectedly easier or harder than their IF ratings anticipated. Items with low IF ratings and high *p*-values were primarily attributed to factoids that were easily recalled despite limited instruction. Items with high IF ratings and low *p*-values were primarily attributed to subtle differences in how the content was taught versus how the content was presented on the examination. In any instance, there are a number of reasons why an item may perform unexpectedly given its IF rating. Instructors are encouraged to investigate the reason for any discrepancy and consider altering their instruction or assessment method appropriately.

### Implications

We believe there are many potential implications for the methodology presented in this paper. First, the methodology is very inexpensive and practical. The only real expense is some additional time to conduct this type of analysis. Most medical educators could perform similar analyses, particularly the elements based on the classical test theory framework, without a sophisticated level of psychometric or statistical knowledge. Next, the methodology promises a great deal of utility with regard to the discernment of authentic and artificial evidence of student learning. The methodology particularly targets the effects of IF on a set of test scores. Currently, there is no other pervasive psychometric methodology that attempts to understand this important factor. While one may never truly know the extent to which a test score accurately reflects what an examinee knows, understanding how IF effects may impact any set of scores is a significant step in the right direction for informing this judgment. Finally, the methodology has significant potential to serve as another source of evidence for the construct validity of test scores and resulting score inferences. In particular, we believe there is a potential for IF to be a recognized and testable property of construct validity.

### Limitations and future research

There are several notable limitations of this methodology and the present study, some of which segue well into additional avenues for further research. First, instructors will need to qualitatively review each item and provide a rating of IF for each. While not a particularly onerous burden, the process will involve some additional time commitment from instructors. Second, the scale presented in this study was very rudimentary. Although, it adequately served the practical purpose of differentiating instructionally familiar items and content, there remains much room for improvement. Future research might focus on developing improved scales, which may include various dimensions of IF (e.g., presentation of content, manner in which it was assessed, perceptions of the examinee, item type, etc.). Relatedly, the authors of this study conceptualized a number of factors that may contribute to IF, but the presented list is hardly exhaustive. Future research should focus on fine-tuning an operational definition to be more exact.

## Conclusion

The purpose of this study was to introduce and describe a relatively simple and straightforward methodology for discerning the effects of instructionally familiar items and content on examinees' scores. Empirical findings that resulted from the psychometric analysis of a moderate-to-high stakes medical school mid-term examination demonstrated the methodology to be robust and capable of achieving its intended purpose. We believe the methodology presented within this paper has significant implications for the discernment of authentic learning and as a potential source of evidence for construct validity. We encourage other medical educators to use this methodology as a model for conducting similar studies of their own.

### Financial support and sponsorship

Nil.

### Conflicts of interest

There are no conflicts of interest.

## References

1. Haladyna T, Roid G. The role of instructional sensitivity in the empirical review of criterion-referenced test items. J Educ Meas 1981;18:39-53.
2. Brennan RL. A generalized upper-lower item discrimination index. Educ Psychol Meas 1972;32:289-303.
3. Cox RC, Vargas JS. A Comparison of Item-Selection Techniques for Norm Referenced and Criterion Referenced Tests. Paper Presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL; 1966.
4. Helmstadter GC. A Comparison of Traditional Item Analysis Selection Procedures with those Recommended for Tests Designed to Measure Achievement following Performance Oriented Instruction. Paper Presented at the Convention of the American Psychological Association, Honolulu, HI; 1972.
5. Kosecoff JB, Klein SP. Instructional Sensitivity Statistics Appropriate for Objectives-Based Test Items. Paper Presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL; 1974.
6. Popham JW, editor. Indices of adequacy for criterion-reference

test items. In: Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, NJ: Educational Technology Publications; 1971. p. 79-98.

7. Roudabush GE. Item Selection for Criterion-Referenced Tests. Paper Presented at the Annual Conference of the American Educational Research Association, New Orleans, LA; 1974.
8. Popham WJ. Instructional sensitivity on tests: Accountability's dire drawback. Phi Delta Kappan 2007;89:146-50, 155.
9. Polikoff MS. Instructional sensitivity as a psychometric property of assessments. Educ Meas Issues Pract 2010;29:3-14.
10. D'Agostino JV, Welsh ME, Corson NM. Instructional sensitivity of a state standards-based assessment. J Educ Meas 2007;12:1-22.
11. McClung MS. Competency testing: Potential for discrimination. Clgh Rev 1977;11:439-48.
12. Mehrens WA, Phillips SE. Sensitivity of item difficulties to curricular validity. J Educ Meas 1987;24:357-70.
13. Linacre JM. Winsteps®, Computer Software, Version 3.75.1. Beaverton, OR (USA); 2013. Available from: http://www.Winsteps.com. [Last accessed on 2014 Mar 17].
14. Wright BD, Masters GN. Number of person or item strata. Rasch Meas Trans 2002;16:888.
15. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. 1st ed. Copenhagen, Denmark: Denmarks Paedagogiske Institut; 1960. p. 184.
16. Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Meas Trans 1994;8:370.
17. Royal KD. Making meaningful measurement in survey research: A demonstration of the utility of the Rasch model. IR Appl 2010;28:2-16.