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

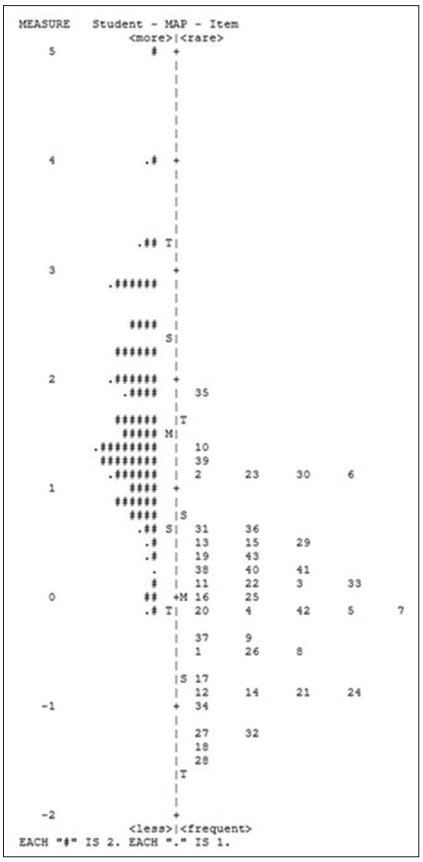

Figure 1: Person and item map

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.