Can Item Keyword Feedback Help Remediate Knowledge Gaps?

Richard A. Feinberg, PhD
Amanda L. Clauser, EdD

ABSTRACT

Background In graduate medical education, assessment results can effectively guide professional development when both assessment and feedback support a formative model. When individuals cannot directly access the test questions and responses, a way of using assessment results formatively is to provide item keyword feedback.

Objective The purpose of the following study was to investigate whether exposure to item keyword feedback aids in learner remediation.

Methods Participants included 319 trainees who completed a medical subspecialty in-training examination (ITE) in 2012 as first-year fellows, and then 1 year later in 2013 as second-year fellows. Performance on 2013 ITE items in which keywords were, or were not, exposed as part of the 2012 ITE score feedback was compared across groups based on the amount of time studying (preparation). For the same items common to both 2012 and 2013 ITEs, response patterns were analyzed to investigate changes in answer selection.

Results Test takers who indicated greater amounts of preparation on the 2013 ITE did not perform better on the items in which keywords were exposed compared to those who were not exposed. The response pattern analysis substantiated overall growth in performance from the 2012 ITE. For items with incorrect responses on both attempts, examinees selected the same option 58% of the time.

Conclusions Results from the current study were unsuccessful in supporting the use of item keywords in aiding remediation. Unfortunately, the results did provide evidence of examinees retaining misinformation.

Introduction

The process of retrieving information from memory has been studied as a means of reinforcing knowledge and facilitating learning.[1,2] Additionally, the process of testing, or forced retrieval of information, may be useful for not just assessment but also knowledge retention.[3,4] Taking a test has been found to improve performance on subsequent tests,[5] with repeated assessment having a positive effect on learning.[6,7] Although the bulk of the cognitive research in this domain has been done in K–12 classrooms and in laboratories, recent research in medical residency has demonstrated that content retention is improved with repeated multiple-choice tests when compared to repeated study (without assessment).[8]

For physicians, multiple-choice examinations are ubiquitous; they appear throughout the United States Medical Licensure Examination sequence and during board certification. These assessments theoretically represent objective measurements of knowledge and are used in conjunction with other processes to determine eligibility for licensure and postlicensure board certification. In anticipation of subspecialty certification examinations, credentialing boards often offer lower stakes in-service or in-training examinations (ITEs) to training programs to better gauge the level of preparedness of fellows. This type of formative feedback has become increasingly common.

The feedback fellows receive can be tremendously important in guiding their study and preparation for future examinations. Receiving feedback of any type is thought to facilitate learning of tested material, "[a]lthough testing improves retention in the absence of feedback . . . providing feedback enhances the benefits of testing by correcting errors and confirming correct responses."[9(p962)] While providing examinees with the correct answers has been found to improve performance and retention,[10] it is not always feasible considering the costs to develop and maintain a secure standardized examination.

When limitations prevent reporting the actual test material, another approach can be to provide item-level keywords. These item-level keywords, sometimes referred to as educational objectives, are brief statements indicating the underlying clinical competency of a particular item and are typically provided to both individual examinees and program directors. For example, the keyword for an item measuring anatomy knowledge could be "femoral nerve block anatomy." Keywords can be provided to examinees for each question responded to incorrectly or, at the program level, all keywords can be provided along

DOI: http://dx.doi.org/10.4300/JGME-D-15-00463.1

with the percentage of trainees who responded correctly. Including keywords in feedback for an ITE is relatively common. Both the American College of Physicians[11] and the American Board of Anesthesiology–American Society of Anesthesiologists[12] report keywords on their ITEs, which are taken each year by approximately 20 000 and 10 000 examinees, respectively.

The current study investigated the utility of providing item-level keywords by assessing whether ITE examinees performed better on items in which the corresponding keywords had been provided as part of feedback from a prior testing attempt.

What was known and gap
Testing, as a forced retrieval of information, may be useful for assessment as well as knowledge.

What is new
This study of 319 fellows, who completed a medical subspecialty in-training examination, assessed whether exposure to item keyword feedback aided in remediation and improved correct response rates.

Limitations
Lack of information about examinees' test preparation; broad keywords used may not provide actionable information to learners.

Bottom line
The use of item keyword feedback was not useful in aiding remediation, with added evidence that examinees retained misinformation.

Methods

Data

Item responses were obtained from fellows who completed a medical subspecialty ITE consisting of 147 multiple-choice items. This ITE is administered annually to approximately 2000 individuals at different points in their postgraduate training. The current study included 319 examinees who tested in 2012 in their first fellowship year and 1 year later in 2013 in their second year, who also responded to the posttest survey question, "How many hours did you spend preparing for this examination?" Of the 319 examinees, 64% selected 0 hours of preparation (n = 204); 17% selected 10 hours (n = 53); 6% selected 20 hours (n = 20); 3% selected 30 hours (n = 8); and 11% selected more than 30 hours (n = 34). Due to the small sample size, the fourth and fifth categories were combined into a group of 42 reporting 30 or more hours of preparation.

After completing the 2012 ITE, examinees received a list of keywords for each item they responded to incorrectly, along with a score report detailing total test and subdomain performance. A fellow who answered 60 items incorrectly would receive a report detailing the 60 content areas and diagnostic/medical terms associated with those items, while a fellow who answered all the items correctly would not receive any keywords. Of the 147 items on the 2013 ITE, 91 were associated with a keyword common to the 2012 form (the same keyword could be linked to more than 1 item). Thus, the content knowledge for 91 of the items on the 2013 ITE was explicitly cued in the 2012 feedback. Of the 91 items that shared a keyword, 27 items were identical on both the 2012 and the 2013 ITEs.

An initial review conducted by the American Institutes for Research Institutional Review Board found this research to be exempt from oversight as it did not involve human subjects and the analyses were based on deidentified data.

Statistical Analysis

To adjust for the possible confounding of differing levels of item difficulty, item response theory using the Rasch model[13] was employed to equate between item sets (eg, exposed versus nonexposed keywords). Scored item response data for second-year fellows testing on the 2013 form was calibrated to produce difficulty estimates for each item and to compute examinee ability on the exposed and nonexposed keyword items. These ability estimates were then converted to scale scores to facilitate interpretation.

To investigate if keyword exposure was related to improved performance on a subsequent testing attempt, a 2×2 mixed design analysis of variance was run with scale scores by item set (exposed or nonexposed keywords) as the within-groups factor and self-reported hours of preparation as the between-groups factor. This includes keywords that may not have been exposed to examinees on their 2012 ITE feedback if they had responded correctly. We followed this approach because (1) restricting the analysis to only keywords presented directly to examinees on their 2012 ITE feedback would have substantially limited the available data, and (2) examinees would have likely been exposed to all keywords on the 2012 form even if they had not responded incorrectly (based on their cohort's performance or other feedback from their program director).

Analyses were conducted on performance patterns for the 27 items common to both forms. Response times and examinee response selection were examined for items that appeared on both the 2012 and 2013 forms.

Results

On the 2013 form, examinees responded correctly to a higher percentage of items in which keywords were

TABLE 1
Performance on the Exposed and Nonexposed Keyword Items for the 2013 Form

Quantity	N	Exposed (n = 91)		Nonexposed (n = 56)	
		Mean	SD	Mean	SD
Rasch item difficulty	147	0.05	1.07	0.25	1.29
Percent correct scores	319	65	9	60	9
Scale scores	319	253	55	247	53
0 hours	204	254	56	248	52
10 hours	53	252	56	245	54
20–30 hours	28	264	60	259	55
More than 30 hours	34	245	48	234	56

Note: Scaled scores are distributed with mean = 250 and SD = 50.

exposed (mean = 65.15; SD = 8.94) than to items in which keywords were not exposed (mean = 59.72; SD = 8.66; $t(318) = 13.55$; $P < .001$; $r = 0.61$). This represents a large effect; however, items associated with nonexposed keywords were more difficult, hence the analysis of scale scores is more appropriate. TABLE 1 shows average performance (scale scores and percentage of items responded to correctly) by examinee self-reported preparation as well as item difficulty for the exposed and nonexposed keyword items.

The analysis of variance revealed a significant main effect of item set on examinee scale score performance ($F_{(1,\ 315)} = 4.08$; $P < .05$; $r = 0.11$). This statistically significant finding indicates that, regardless of preparation, examinees performed better on the exposed keyword items. However, the effect size is very small, accounting for only 1% of the variance. There was no significant main effect of preparation ($F_{(3,\ 315)} = 1.02$; $P = .39$; $r = 0.06$) and no significant interaction effect between item set and preparation ($F_{1315} = 0.23$; $P = .88$; $r = 0.03$), indicating that performance was unrelated to the amount of preparation time reported. These results are illustrated in the FIGURE. Overall, performance was slightly higher on the exposed keyword items; however, the slope of the lines remained consistent across levels of preparation.

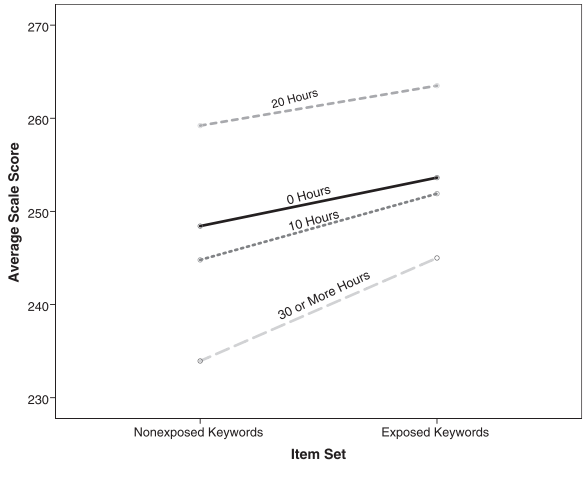

FIGURE
Relationship Between Performance on Exposed and Nonexposed Keyword Items and Self-Reported Examinee Preparation

Analysis of response patterns for the 27 items common to both forms are presented in TABLE 2. In total there were 8613 pairs of responses (319 examinees × 27 items). Examinees responded correctly to a higher percentage of the same items in 2013 (sum of the first and third rows = 64%) than in 2012 (sum of the first and second rows = 52%), substantiating the theory that their performance improved (based on overall growth) between training years. The second row represents examinees who may have "guessed lucky" on their first attempt, although the increase in response time (approximately 14 seconds on average) suggests that they may have forgotten the content and spent extra time unsuccessfully trying to remember.

Of the 48% incorrect responses on the 2012 attempt, we would expect, given that each question had at least 5 options, that 20% or 9.6% of incorrect responses to be converted to correct by chance alone. The result that 23% of responses moved from incorrect to correct demonstrates that some learning took place. Of the 25% responses that were incorrect on both attempts (n = 2173), examinees selected the same incorrect option about 58% of the time. TABLE 3 presents the cross-tabulation for each pair of options. For instance, of the examinees who selected "A" on the 2012 ITE (where "A" was not the correct option but a distractor), 60% selected "A" to the same question during the 2013 ITE.

Discussion

The main findings from this study were that examinee preparation was unrelated to performance

TABLE 2
Common Item Response Patterns and Change in Response Time (N = 8613)

Same Item Response Pattern (Year 1 Response – Year 2 Response)	%	Change in Response Time (Seconds)	
		Mean	SD
Correct–Correct	41.4	–4.91	63.27
Correct–Incorrect	10.4	14.04	67.69
Incorrect–Correct	23.0	3.32	70.40
Incorrect–Incorrect	25.2	8.39	71.01

Note: Changes in response time were calculated as 2013 duration minus 2012 duration.

differences between items with exposed and nonexposed keywords, and that for common items, examinees who responded incorrectly on the 2012 ITE selected the same incorrect response option 1 year later more than half the time.

Without a significant interaction between preparation and improvement on items in which the keywords were exposed, performance differences between the 2012 and 2013 ITEs are likely due to other factors. For instance, concepts that reappear on future versions of the test are likely highly relevant to the fellowship curriculum, and thus performance on the exposed keywords may have increased based solely on additional medical training.

The high probability of selecting the same incorrect response option suggests these examinees were misinformed, and this error influenced both administrations. Furthermore, it is possible that without immediately correcting the error, examinees may have acquired false knowledge by believing their original response was correct.[5] If keywords can help examinees identify misinformation, then we would expect that the probability of selecting the same incorrect response option would have been closer to chance.

A limitation of any study on keyword feedback is that they are based on single-item responses and may not provide reliable information about what examinees do or do not know. Additionally, for this study, we do not know *how* time was spent in preparation. Examinees were not explicitly directed on how to prepare for the 2013 ITE and the reported preparation time may have been used to study other materials or resources, ignoring the keyword feedback from the previous test. Another limitation is that the keywords for this ITE tend to be broad (eg, coordinate patient care and handoffs, including transition or transfer of care), and may have lacked sufficient specificity to help examinees target their knowledge deficits.

Future research can help further understanding of the utility of providing keywords by investigating how examinees use keywords to identify and remediate knowledge deficits. Additionally, future research should also explore keyword specificity to determine how they can best assist examinees in identifying misinformation in the absence of providing test material.

Conclusion

The results were unsuccessful in supporting the use of item keywords in aiding remediation. Results also provided evidence of examinees repeating errors from year-to-year, which suggests that, without sufficient remediation, errors may go uncorrected. Further exploration is warranted to determine if the

TABLE 3
Answer Option Selection for Items Scored Incorrect on Both Forms (N = 2173)

2012 ITE Response	2013 ITE Response (%)					
	A	B	C	D	E	F
A	**60.0**	10.3	13.0	9.6	6.5	0.7
B	12.0	**48.3**	13.9	8.0	16.0	1.6
C	12.0	9.5	**61.9**	6.2	8.3	1.8
D	16.1	12.1	7.1	**54.0**	10.2	0.0
E	6.8	10.4	10.9	7.3	**63.5**	1.1
F	4.9	4.9	17.1	0.0	14.6	**58.5**

Abbreviation: ITE, in-training examination.
Note: Values in boldface indicate the percentage of examinees selecting the same incorrect response option on both attempts.

lack of validity for keywords demonstrated in this study is, in fact, due to the keyword feedback itself or to the way(s) examinees used (or did not use) feedback to prepare for their second attempt at the examination.

References

1. Rock I. The role of repetition in associative learning. *Am J Psychol*. 1957;70(2):186–193.
2. Loftus EF, Pickrell JE. The formation of false memories. *Psychiatr Ann*. 1995;25:720–725.
3. Wheeler M, Roediger HL. Disparate effects of repeated testing: reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychol Sci*. 1992;3(4):240–245.
4. Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci*. 2006;17(3):249–255.
5. Bangert-Drowns RL, Kulik JA, Kulik CC. Effects of frequent classroom testing. *J Educ Res*. 1991;85(2):89–99.
6. Marsh EJ, Roediger HL, Bjork RA, et al. The memorial consequences of multiple-choice testing. *Psychon Bull Rev*. 2007;14(2):194–199.
7. Roediger HL III, Marsh EJ. The positive and negative consequence of multiple-choice testing. *J Exp Psychol Learn Mem Cogn*. 2005;31(5):1155–1159.
8. Larsen DP, Butler AC, Roediger HL 3rd. Repeated testing improves long-term retention relative to repeated study: a randomized controlled trial. *Med Educ*. 2009;43(2):1174–1181.
9. Larsen DP, Butler AC, Roediger HL 3rd. Test-enhanced learning in medical education. *Med Educ*. 2008;42(10):959–966.
10. Pashler H, Cepeda NJ, Wixted JT, et al. When does feedback facilitate learning of words? *J Exp Psychol Learn Mem Cogn*. 2005;31(1):3–8.
11. American College of Physicians. Internal medicine in-training examination. https://www.acponline.org/featured-products/medical-educator-resources/im-ite. Accessed April 18, 2016.
12. American Board of Anesthesiology. In-training examination. http://www.theaba.org/TRAINING-PROGRAMS/In-training-Exam/About-the-In-Training-Exam. Accessed April 18, 2016.
13. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.

Both authors are at the National Board of Medical Examiners. Both **Richard A. Feinberg, PhD,** and **Amanda L. Clauser, EdD,** are Psychometricians.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

Corresponding author: Richard A. Feinberg, PhD, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, 215.590.9553, rfeinberg@nbme.org

Received September 15, 2015; revisions received December 30, 2015, and February 24, 2016; accepted March 1, 2016.