Assessing the viability of External Searchable Resources on the American Board of Family Medicine's certification examination

Thomas R. O'Neill, Ph.D.

Michael R. Peabody, Ph.D.

American Board of Family Medicine

Keith L. Stelter, MD, MMM

University of Minnesota - Mankato

Michael D. Hagen, MD

American Board of Family Medicine; University of Kentucky

Publication Date: 7/18/2015

INTRODUCTION

The advent of personal computing devices such as handheld tablet computers and smart phones with high-speed Internet connectivity has placed a myriad of medical information sources at the physician's disposal at the point of care. In addition to the prompts that are built into Electronic Health Record (EHR) interfaces, physicians have access to subscription-based clinical resources such as UpToDate, DynaMed, Epocrates, ACP Smart Medicine, and Isabel, and these subscription-based resources may either stand alone or integrate with the EHR interface. A recent survey by Wolters Kluwer Health (2011) found that 46% of physicians cited Internet search engines such as Google and Yahoo as frequent sources of information; 42% indicated frequent use of free online services (e.g. WebMD, MayoClinic.com); and 36% identified frequent use of online subscription services (e.g. UpToDate, EBSCO).

Although the availability of external searchable resources (ESR) has increased dramatically and their use is becoming more common, the extent to which they are being used at the point-of-care and the type of information being researched (dose calculator or vaccine schedule vs diagnostic and therapeutic information) is not clearly understood. Active prompts from an EHR are only presented at the point-of-care; however, these usually consist of basic features like prompts for preventive health maintenance issues such as vaccines and colorectal cancer screening or ACE-inhibitor use in patients with diabetes. On the other hand, physicians can use subscription-based searchable resources and free searchable resources either at the point of care or at other times to research current best practices. The increasing prevalence of these ESRs in practice has prompted a few calls to consider incorporating such resources into board certification examinations (Cooke, 2013)

The role of the examination

The American Board of Medical Specialties (ABMS) requires member boards to implement a four-part maintenance of certification (MOC) process; however, the specific processes are developed by each individual member board. ABMS boards intend their certification to be a public attestation that a physician has: adequate professional standing [Part I], an ongoing commitment to lifelong learning [Part II], sufficient clinical decision-making ability supported by a suitable fund of medical information [Part III], and a commitment to practice improvement [Part IV] (Rinaldo & O'Neill, 2009; Hawkins, Lipner, Ham, Wagner, & Holmboe, 2013; Nora, 2013; Lipner, Hess, & Phillips, 2013).

The American Board of Family Medicine (ABFM) meets the ABMS Part III requirement by administering the Maintenance of Certification for Family Physicians (MC-FP) Examination. The MC-FP Examination is a secure, closed-book, proctored, standardized, summative examination. It yields a pass-fail decision intended to reflect whether a physician can sufficiently utilize current best practices in their clinical decision-making across the entire spectrum of family medicine content (Norris, Rovinelli, Puffer, Rinaldo, & Price, 2005). The MC-FP Examination is high stakes in that the test is only offered twice a year and failing physicians, who did not allow for sufficient time to retest before their certification expires, will experience a gap in their certification. Test questions reflect commonly encountered clinical situations in Family Medicine as well as less common diagnoses that are important to recognize. These questions undergo extensive editing and review to exclude minimally relevant information. The ABFM's certification process reflects the broad-spectrum of family medicine, not an individual's personal scope of practice (O'Neill, Peabody, Blackburn, & Peterson, 2014).

Assessing the Need for ESRs on the MC-FP Exam

ABFM certification is not conditional upon the resources that the physician has available. Indeed, what medical knowledge a physician should hold in memory versus what may be reasonably relegated to an electronic repository remains an open question, but evidence does exist that successful practice requires that physicians maintain a substantial and critical core of information in their personal knowledge base (Holmboe, Lipner, & Greiner, 2008).

The item review process specifically considers whether candidates should be able to answer the questions using only the physician's personal fund of knowledge. During the review, items that do not meet this criterion are eliminated from the examination.

This study examined whether *physician raters* felt that the items on the MC-FP Examination would require the use of an ESR and the perceived extent to which an ESR was needed to answer individual questions. It should be noted that this differs from asking if candidates *wanted* to use an ESR during the examination. More resources are usually desirable, but this study investigated whether the participants felt they were necessary.

METHODS

Participants

In order to conduct the study, we integrated it into the standard-setting process for the MC-FP Examination. For our standard-setting process, we invite a group of volunteer Diplomates who passed the exam in the previous year to rate a sample set of items from the examination using an Internet-based application. The initial call for volunteers was sent to 2,395 Diplomates and within a few days 287 had accepted the offer; 144 were subsequently selected to participate in the standard setting study. Of these 144 volunteers, 122 completed the required web-based training sessions, 91 completed the standard setting process, and 87 provided usable

data. Table 1 shows the representativeness of the relevant demographic characteristics of the sample of raters compared to that of the population from which they were selected.

Instrumentation

The modified-Angoff method (Angoff, 1971) is a widely-used, content-based method for recommending a passing standard (Hurtz & Auerbach, 2003; Plake & Cizek, 2012). This method asks content experts to examine a representative set of test questions and for each question determine the probability that a "minimally acceptable examinee" would answer it correctly. This probability rating is commonly referred to as a p-value and ranges from .00 to 1.00, with a higher rating indicating a greater probability of an examinee getting a question correct. The ABFM's web-based application for collecting these ratings was designed so that the rater must answer the question under review and provide an initial p-value rating for that question. Next, the rater is provided feedback in the form of the correct answer to the question and the probability that an examinee with a score at the current passing would correctly answer the question. After considering the feedback, raters could elect to revise their initial rating when entering their final rating. This procedure encourages a thoughtful consideration of each item prior to asking the rater whether the item required an ESR to answer it. The response options for this question were:

1. They should not need to look this up
2. A few might need to look this up
3. Most would need to look this up
4. They should look this up often

In this study, 120 test items were selected to mirror the test plan specifications for the MC-FP Examination and each rater was asked to rate the same 120 test items. It should be noted that the

ABFM's web-based application differs from more traditional versions of the modified-Angoff method in that it is conducted asynchronously and raters do not discuss the items face-to-face as is typical in many studies that use the modified-Angoff method.

Data Analysis

To ensure that the results can be interpreted meaningfully, the functioning of the ESR rating scale (reliability and fit of data to the model) was examined using the Rasch Rating Scale Model (Andrich, 1978). To make sense of the ratings, it may be helpful to interpret the mean ratings for both raters and items by assigning the means to the nearest rating scale category. For each of the 120 items, both the mean Angoff rating and mean ESR rating was computed. Pearson correlation coefficients were calculated in order to determine the strength of the association between the rater's perception of item easiness (using their Angoff-derived rating) and their ESR rating. Finally, two scalograms (Guttman, 1944) were constructed by ordering the raters top-to-bottom based on their willingness to endorse items and items left-to-right by the easiest to hardest to endorse items. This map initially included all responses and then was repeated using only responses for questions that they answer correctly.

RESULTS

Functioning of the ESR Rating Scale

The reliability of the raters, items, and the rating scale were estimated. The items produced a Rasch person reliability (Smith, 2000; Linacre, 2002) of .95. The mean ESR person rating was 2.2 (SD=.3). The average person infit and outfit mean-square values were both 1.00 (Table 2). The people produced a Rasch item reliability of .97. The mean ESR item rating was 2.2 (SD=.4).

In addition to the Rasch Rating Scale summary statistics, Table 2 also shows the counts and percentage of responses for each rating scale category. Each of the 87 raters provided ratings on 120 items for a total of 10,440 responses: category 1 had 20% of the responses (N=2,136), category 2 had 42% of the responses (N=4,407), category 3 had 32% of the responses (N=3,336), and category 4 had 5% of the responses (N=561). Rating scales with Andrich thresholds that monotonically advance by more than 1.2 logits per category (assuming 4 categories) indicate that the question functions similarly to 3 dichotomous items (Linacre, 2002). The increases in Andrich thresholds on this rating scale exceeded this criterion. The mean observed category values also increased monotonically, which indicates that this rating scale meets the requirements of measurement and provides more information per item than a dichotomous item.

Table 3 shows the results when the mean ratings for both raters and items are assigned to the nearest rating scale category. Of the 87 raters, 78 (90%) had a mean rating that would fall into category 2, *a few might need to look this up*. The remaining 9 (10%) had a mean rating that would fall into category 3, *most would need to look this up*. The distribution of mean ratings (Figure 3) suggests that 3 or 4 of the 9 raters classified as category 3 might be outliers. Of the 120 items, 88 (73%) had a mean rating that would fall into category 2, *a few might need to look this up*.

Correlation

Rater perception of item easiness was negatively correlated with their perception that an ESR was needed for the item (r = -.94). Figure 1 provides a plot of the items comparing their mean Angoff rating with their mean ESR rating.

Scalograms

The scalograms were based on responses to the ESR question, not on the medical ability of the rater or the difficulty of the item; although, those issues might be related to the ESR rating. Due to the size of the scalogram in this study, we transformed the responses onto a grayscale to assist with visibility. The higher the rating the darker the grayscale, so that the highest rating of "*They should look this up often*" is the darkest shade and the lowest rating of "*They should not need to look this up*" is the lightest shade. For this study, we have constructed two scalograms: the scalogram at the top of Figure 2 represents all of the responses, while the scalogram below it represents only responses from raters who answered the question correctly. In the second scalogram, incorrect responses are indicated with a white box.

DISCUSSION

Perceived Need for an ESR

This study examined whether raters felt that the items on the MC-FP Examination would require the use of an ESR and the perceived extent to which an ESR was needed to answer the question. This differs from asking if candidates wanted an ESR on the examination. More resources are usually desirable, but this study investigated whether the participants felt they were needed.

Using the categorization schema from Table 3, each of the 120 items were classified by their mean rating. There was general consensus among raters that very few items should never be looked up and no items should always be looked up. Most of the items (73%) were rated as, *a few might need to look this up*. As previously noted, the MC-FP Examination questions are designed in such a way that an ESR should not be needed to answer them correctly. The item review process specifically considers whether candidates should be able to answer the questions

using only the physician's personal fund of knowledge. During the review, items that do not meet this criterion are eliminated from the examination. The results from this survey suggest that these efforts have been largely successful.

Issues Related to Implementing an ESR

While this study suggested that an ESR is currently not necessary to pass the MC-FP exam, the question remains open as to whether the exam *should* include questions requiring ESRs. Making an ESR available would likely be popular with examinees because it would seem to increase the chances of an examinee correctly answering the question. The examinees might also perceive this as a more congruent reflection how they practice. However, allowing ESRs in the MC-FP Examination would require several issues to be resolved. First, what is the MC-FP Examination intended to measure? Are efficient search skills that produce accurate results part of the intended evaluation construct? The examination is currently constructed as a test of clinical decision-making ability supported by a sufficient fund of medical information, but the implementation of ESRs would overlay on this model an assessment of the physician's ability to search for information efficiently and accurately. In this scenario, every item could be about clinical decision making and search efficiency and search accuracy. When someone gets a question wrong, it might be because (1) they were confident in the wrong answer, (2) perhaps they knew the answer but it conflicted with their search results and they answered based on the search result, or (3) they did not know the answer and their search skills were poor.

If an examinee's concern is passing the examination rather than mirroring their manner of practice, then making an ESR available could likely have an unintended consequence. If an ESR was used on difficult items and it made a difference in examinees' ability to correctly answer them, those items would become easier. When there are many easy items on an examination,

examinees would have to answer more items correctly to pass. Adding an ESR could increase the number of items that people must answer correctly to pass. In this study, the perceived need for an ESR was strongly correlated with item difficulty (Figure 2) which suggests that using an ESR would likely make the items easier.

Making an ESR available would likely also result in another unintended consequence. Examinees may feel that they should research each question because it seems indefensible to get questions wrong on what seems to be the equivalent of an open-book test. This would increase the time demands of the test. If extra time is not made available, the test sponsor will be subject to a different potential challenge from examinees: the examination is a test of speed rather than clinical decision-making and knowledge. Furthermore, examinees who receive an extended-time accommodation on the exam would have an advantage over those candidates who do not receive such extra time. If extra time was given to all candidates, the examination would be slightly more expensive due to the increase in test center time, but more importantly the examination could go into a second day which could be a burden on a physician's practice.

There are other obstacles that must be overcome before one could implement an ESR, such as: which ESRs would be allowed, how to ensure the security of items, and how to deal with search results that conflict among ESRs and conflict with the citation/s that the certification board uses to justify the correct answer. Using one ESR but not others might make the board vulnerable to accusations of favoring one product or vendor over another. Physicians who use a different ESR in their practice might be at a disadvantage using an unfamiliar product. Conversely, making the entire internet available to examinees could present a serious security problem. Item harvesters would find it easy to send the general content of the items to an external repository. It could also make the ESR function similar to the "phone-a-friend" option if

they could "chat" with an expert physician who could also look things up. Finally, item writers go to great lengths to research and document the correct answers to questions with reproducible evidence based research and search results may differ between ESRs simply based on the algorithm used by developers. This could provide examinees with an unwarranted sense of confidence regarding the correctness of their response.

Limitations

For each item, the raters were told whether or not they had correctly answered the question and were provided the empirical item difficulty before being asked to rate the item for whether an ESR was needed. This may have altered the raters' perception of whether the item should be looked up. For this reason, the scalogram was presented first with all responses and second with the ESR rating removed if the rater answered the question incorrectly (Figure 2). It seems that the many of incorrect responses were items that were rated 3 or 4, suggesting that raters tended to want to look up items they were told they answered incorrectly.

All of the raters participating in this study had recently passed the exam; no failing candidates were asked to participate. On one hand, these raters will not have to take the examination for another 9 or 10 years, so they might be reasonably free from test-related anxiety and their response may reflect more of their thoughts than feelings about the issue. However, using this sampling strategy effectively excludes a small sample of the testing population, the lower performers, from this study.

CONCLUSION

The decision to implement an ESR should be based upon whether such access is beneficial or at least neutral with regard to the certifying boards' mission of assuring the public regarding certified Diplomates' abilities, and whether the obstacles to implementation can be

overcome without harming the integrity of the examination. Although this study used ABFM items and physicians, and the results are specific to ABFM, the issues are more global in scope. Allowing examinees the ability to search references would alter the construct of the Part III examination and place an excessive emphasis on an examinee's ability to efficiently and accurately search an ESR. Furthermore, a physician's certification exam score could be unfairly linked to their ability to navigate a specific ESR.

For the ABFM, and many other certification organizations, test questions usually cannot be solved using only rote memorization. Although some questions could be considered to be in the "this is treated with that" format, many of the vignettes are designed to be more complex. Physicians, and family physicians in particular, often deal with complex patients who have multi-morbid conditions. These are not easily researched and take a certain level of expertise to understand. These questions are difficult to write and then document with an empirically supported correct answer; however, ABFM works hard to do just that.

If certifying organizations consider information retrieval skills as important, it might be more appropriate to address this in another aspect of the maintenance-of-certification process, such as in Part II and/or Part IV. For the ABFM, Part II serves to update and enhance the physician's knowledge base of clinical best practices; in that venue the use of an ESR is not only appropriate, but desirable. Part III serves a different purpose. It functions not as a means to help physicians update their knowledge-base, but rather to demonstrate that they have in fact successfully updated and maintained their medical knowledge.

The purpose of ABFM certification is to publicly attest using a standardized process that a physician has adequate professional standing, a high level of medical knowledge, and an ongoing commitment to lifelong learning and performance improvement, so that the potential

patient or provider can make a judgment whether to trust a particular physician.[3] This independent attestation regarding physician quality might be perceived differently by the public if an ESR were permitted. If the public felt that the test was essentially an open-book examination, would they feel that the medical profession is doing a sufficient job of regulating itself? We believe that passing a rigorous, secure, closed-book, standardized, summative examination is crucial assuring the public of certified family physicians' clinical acumen.

REFERENCES

Andrich, D. (1978). A Rating Formulation for Ordered Response Categories. *Psychometrika, 43*(4), 561-573.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508-600). Washington, DC: American Council on Education.

Cooke, M. (2013, October 8). President's Message: Maintenance of Certification is needed, but it needs to change. *ACP Internist*.

Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A., & Bogels, S. M. (2010). The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. *Sleep Med Rev, 14*(3), 179-189.

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review, 9*(2), 139-150.

Hawkins, R. E., Lipner, R. S., Ham, H. P., Wagner, R., & Holmboe, E. S. (2013). American Board of Medical Specialties Maintenance of Certification: Theory and Evidence Regarding the Current Framework. *Journal of Continuing Education in the Health Professions, 33*(S1), S7-S19.

Hurtz, G. M., & Auerbach, M. A. (2003). A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus. *Educational & Psychological Measurement, 63*(4), 584-601.

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Lipner, R. S., Hess, B. J., & Phillips, R. L. (2013). Specialty Board Certification in the United States: Issues and Evidence. *Journal of Continuing Education in the Health Professions, 33*(S1), S20-S35.

Nora, L. M. (2013). Professionalism, Career-Long Assessment, and the American Board of Medical Specialties' Maintenance of Certification: An Introduction to This Special Supplement. *Journal of Continuing Education in the Health Professions, 33*(S1), S5-S6.

Norris, T. E., Rovinelli, R. J., Puffer, J. C., Rinaldo, J., & Price, D. W. (2005). From Specialty-Based to Practice-Based: A New Blueprint for the American Board of Family Medicine Cognitive Examination. *Jounal of the American Board of Family Practice, 18*(6), 546-554.

O'Neill, T. R., Peabody, M. R., Blackburn, B. E., & Peterson, L. E. (2014). Creating the Individual Scope of Practice (I-SOP) Scale. *Journal of Applied Measurement, 15*(3), 227-239.

Plake, B. S., & Cizek, G. J. (2012). Variations on a Theme: The Modified Angoff, Extended Angoff, and Yes/No Standard Setting Methods. In G. J. Cizek (Ed.), *Setting Performance Standards* (2nd ed., pp. 181-199). New York: Routledge.

Rinaldo, J. C., & O'Neill, T. R. (2009). The Measure of Family Medicine - Response 3. *Family Medicine, 41*(8), 539-540.

Smith, R. M. (2000). Fit Analysis in Latent Trait Measurement Models. *Journal of Applied Measurement, 1*(2), 199-218.

Wolters Kluwer Health. (2011). Point of Care Survey.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions, 8*(3), 370.

TABLE 1
Rater Demographic Information

	% of Sample	% of Population
Gender		
Male	44%	58%
Female	56%	42%
Certification Status		
Initial Certification	15%	30%
Recertification	85%	70%
Medical Degree		
M.D.	87%	90%
D.O.	13%	10%
Medical School Training		
U.S.	86%	81%
Canada	1%	1%
Other International	13%	18%

Note. The population was drawn from those who passed the Spring exam administration.

TABLE 2
Rating Scale Summary Statistics

	Mean	Infit		Outfit		
	Measure	MNSQ	ZSTD	MNSQ	ZSTD	Reliability
Rater mean (SD)	-0.71 (.67)	1.00 (.38)	-.4 (3.2)	1.00 (.38)	-.4 (3.2)	.95
Item mean (SD)	0.00 (1.04)	.99 (.23)	-.1 (1.6)	1.00 (.23)	-.1 (1.6)	.97

Rating Scale Categories

Label	Count (%)	Obs. Avg.	Infit MNSQ	Outfit MNSQ	Andrich Threshold
1. They should not need to look this up.	2,136 (20%)	-1.76	1.05	1.05	NONE
2. A few might need to look this up.	4,407 (42%)	-1.00	0.90	0.90	-2.11
3. Most would need to look this up.	3,336 (32%)	0.09	0.91	0.92	-0.19
4. They should look this up often.	561 (5%)	0.90	1.14	1.15	2.30

TABLE 3
Criteria for assigning mean ESR ratings to rating scale categories

Rating Scale Category	Ranges	Mean Person Rating		Mean Item Rating	
		N	%	N	%
1. They should not need to look this up.	1.0 – 1.5	0	-	4	3%
2. A few might need to look this up.	1.6 – 2.5	78	90%	88	73%
3. Most would need to look this up.	2.6 – 3.5	9	10%	28	23%
4. They should look this up often.	3.6 – 4.0	0	-	0	-

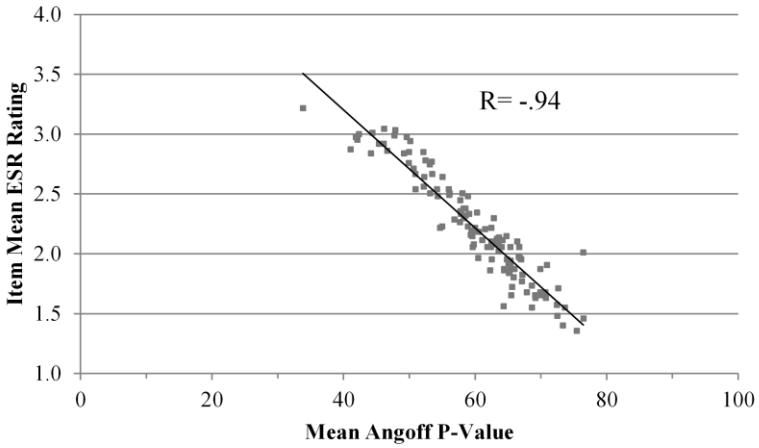

FIGURE 1
Plot of Angoff-derived item ratings with ESR ratings

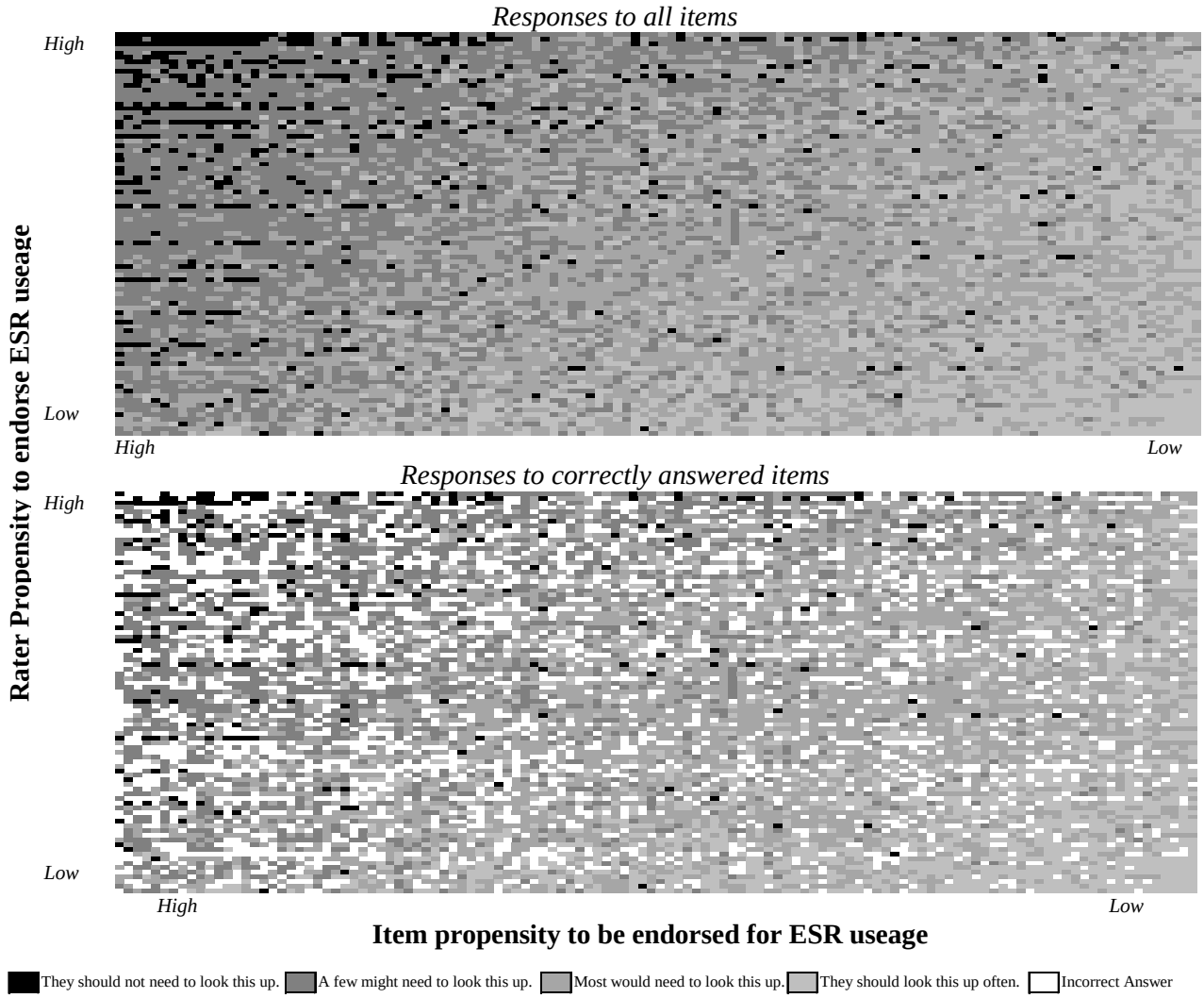

FIGURE 2

Comparison of Scalograms for all responses and the responses to correctly answered questions

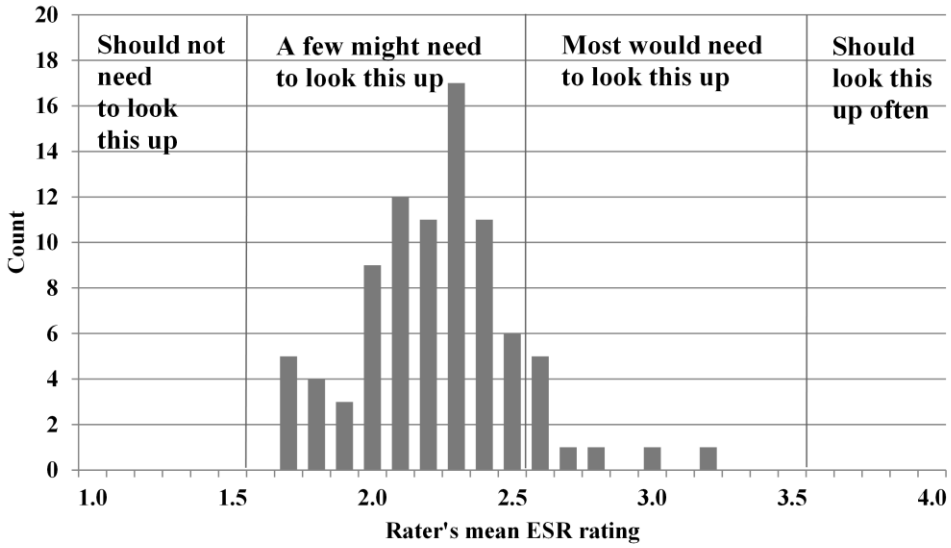

FIGURE 3
Histogram of Raters' mean ESR ratings across all 87 raters