Practical Assessment, Research, and Evaluation

Volume 24 *2019* Article 2

2019

Determining Item Screening Criteria Using Cost-Benefit Analysis

Bozhidar M. Bashkov

Jerome C. Clauser

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 24 Number 2, April 2019 ISSN 1531-7714

Determining Item Screening Criteria Using Cost-Benefit Analysis

Bozhidar M. Bashkov, *American Board of Internal Medicine*
Jerome C. Clauser, *American Board of Internal Medicine*

Successful testing programs rely on high-quality test items to produce reliable scores and defensible exams. However, determining what statistical screening criteria are most appropriate to support these goals can be daunting. This study describes and demonstrates cost-benefit analysis as an empirical approach to determining appropriate screening criteria for a given testing program and purpose. Using a certification exam's item pool and simulation we illustrate how to examine a wide range of screening criteria and reach an acceptable balance between the number of items screened out (cost) and pass/fail classification accuracy (benefit).

Educational testing practitioners are well aware that successful testing programs rely on high-quality test items. That is, in addition to covering the breadth of the construct being measured as outlined in the test blueprint, items need to be statistically acceptable to be included on a test. What measurement professionals do not necessarily agree upon is which items can be considered "good enough." What specific criteria do test items need to meet to be deemed psychometrically sound? To answer this question, psychometricians often defer to item screening rules, which essentially specify the range of values item difficulty and discrimination can take to be deemed acceptable. These rules are typically enforced at pretesting of the items, and the items' performance is monitored throughout their lifetimes as live operational items. The use of classical screening rules may seem obsolete in a world where large-scale testing programs increasing rely on Item Response Theory (IRT); however, as will become evident in the next sections, many practitioners, including large testing programs like the SAT, continue to use classical statistics to screen items at pretest. The use of classical statistics is particularly common among testing programs with small examinee samples, where it may not be feasible to obtain stable IRT parameter estimates and information (Zumbo & Rupp, 2004, Chapter 4). Despite the continued reliance on screening rules using classical statistics, there is considerable disagreement in the literature as to what these screening rules should be. In this paper, we review the literature on classical screening criteria to highlight the variety of rules recommended in different contexts, and then we describe and demonstrate the use of cost-benefit analysis, a framework commonly used in economics, to determine an appropriate set of screening criteria for a given testing program and purpose empirically.

Guidelines from the Literature

The Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) do not provide any specific screening criteria but state that the model and the sample used for evaluating the psychometric properties of items should be justified and well documented (Standard 4.10, p. 88). Nevertheless, item screening guidelines do exist. More than five decades ago, Ebel (1965) proposed what are perhaps the earliest item screening criteria. His guidelines pertained to the so-called index of discrimination D (i.e., the difference in an item's proportion correct or p-values between the top and bottom 27% of examinees based on their total scores) ranging from -1 to 1. The use of 27% is somewhat arbitrary but can be traced back to Kelley (1939). Ebel's guidelines suggested that items with $D <$

.20 needed to be heavily revised or eliminated; items with *D* of .20-.29 were marginal and needed revision; items with *D* of .30-.39 needed little or no revision; and items with *D* of .40 or higher were functioning well. However, these recommendations were based on anecdotal experience rather than empirical evidence. Although the index of discrimination has fallen out of favor (Youngman, 1979), the precedent of relying on experience rather than empirical analysis to establish screening criteria has persisted.

Discrimination

Most screening recommendations available today differ based on both the context of the exam and the author offering the recommendation. For example, Varma (2010) in a recommendation for school practitioners, endorsed an item-total point-biserial correlation of 0.15, but suggested that "[her] experience has shown that 'good' items have point-biserials above 0.25." (p. 6). In a somewhat more lenient recommendation, the Iteman user manual (Assessment Systems Corporation, 2017) recommends that screening rules be set at 0.10 or 0.20, but suggests that this cutoff could be reduced further in the event of small sample sizes. Clauser and Hambleton (2017) are more lenient still, indicating that in the context of classroom assessments "the expectation is that the discrimination values are at least above 0.0 and preferably substantially so" (p. 360). Although partly justified by context, this discordance across recommendations is enough to stymie practitioners, but it is not the end of the confusion in the interpretation of screening rules.

Even when it appears that recommended screening rules agree, discrepancies exist due to the chosen correlation statistic. The most common issue seems to be between the biserial correlation and the point-biserial correlation. Both correlation metrics indicate the relationship between one continuous and one dichotomous variable; however, the biserial assumes an underlying continuous variable (e.g., ability) manifested in a dichotomous one (e.g., correct/incorrect), whereas the point-biserial does not have this assumption. These are by far the most widely used item discrimination indices in measurement practice today, but as they are underpinned by different assumptions and computed differently, they result in different values for the same data. Thus, even when screening recommendations appear to be nominally the same, they have practically different interpretations. For example, in 1967 Nunnally recommended a minimum biserial correlation of 0.20. More recently, Nunnally and Bernstein (1994) again recommended a minimum discrimination of 0.20, but now this recommendation was applied to the point-biserial correlation (pp. 302-306).

In fact, the recommendation of 0.20 appears many times in the literature for both the biserial and the point-biserial. For example, ACT (Liu, Harris, & Schmidt, 2007) and Educational Testing Service (Chubbuck, Curley, & King, 2016) report 0.20 as the minimum recommended biserial correlation. Similarly, the technical manual for the SAT explains that "most items included on SAT operational forms fall within a biserial range of +.20 to +.80" (The College Board, 2016, p. 46). Crocker and Algina (1986), on the other hand, make the recommendation of 0.20 as the minimum point-biserial (p. 324). Finally, in the fourth edition of *Educational Measurement*, Schmeiser and Welch (2006) state that "in a test intended to rank order and differentiate examinees, discrimination indices of .20 are desirable" (p. 339) but do not specify whether this recommendation applies to the biserial or point-biserial correlation. This may seem like a minor point, but since the biserial correlation is almost always higher (or more extreme in an absolute sense) than the point-biserial correlation for the same data (Lord & Novick, 1968), the recommendations for the same value can differ substantially. For example, a point-biserial of 0.20 for a mid-difficulty item corresponds to a biserial of 0.25 (Terrell, 1982). A difference of this magnitude will have a significant impact on item selection decisions for many exams. Overall, although the 0.20 minimum has been a recommended screening rule repeatedly, there is little consensus in the literature since this recommendation has been applied to at least three different correlation indices.

In sum, there are substantial differences across the literature in what constitutes a minimally acceptable level of discrimination. Nominally, the differences seem to range from 0.0 to 0.25, but these differences are exacerbated by the choice of correlation statistics. This discordance places the measurement professional in a difficult position when attempting to select an appropriate discrimination screening criterion for their particular exam.

Difficulty

Guidelines for item difficulty are less common than those for discrimination, as many authors suggest that decisions of item difficulty criteria be driven predominantly by test purpose. For example, Schmeiser and Welch (2006) warn that extremely difficult items ($p < .30$) could be problematic, especially when discrimination is negative, but acknowledge that determining this lower bound will depend on the test's purpose. Crocker and Algina (1986), reaffirming the focus on test purpose, explain that when items are moderately correlated with the total score or an external criterion (e.g., r_{bis} = 0.30-0.40), mid-difficulty items (e.g., p = 0.40-0.60) are preferred as they tend to maximize reliability of scores across a wide range of proficiency. However, for a criterion-referenced test whose goal is to select among the best applicants, it may be desirable to have more difficult items to ensure sufficient discrimination among examinees at this range of ability. With respect to the other end of the difficulty spectrum, authors also advise selecting cutoffs based on the test's context and purpose with *p*-values as high as 0.95 considered acceptable in situations where most examinees are expected to know the content such as end-of-course exams (Schmeiser & Welch, 2006). In general, the recommendations for item difficulty tend to be more consistent than those for item discrimination, but they are also much less prescriptive. These recommendations create ambiguity as to what exactly one should do operationally. In other words, although there is a general consensus as to *what* practitioners should do (i.e., determine screening criteria based on the test's purpose and the examinee population), the literature provides little guidance on *how* practitioners should go about performing this task.

In summary, there is significant variability in the recommended screening criteria found in the literature. This would be a relatively small problem if there were substantial empirical evidence to support these recommendations. Unfortunately, to our knowledge, no citations are provided for these recommended guidelines, so it remains unknown how adopting a higher or a lower criterion would impact important psychometric outcomes such as examinee scores, classification decisions, and ultimately the validity of the inferences made about examinees (Kane, 2013). This is regrettable since, as we will demonstrate, establishing screening criteria based on empirical evidence makes it possible to identify an appropriate solution for a testing program given its examinee population, item characteristics, and purpose.

Determining Item Screening Criteria Empirically

The purpose of this study is to illustrate an empirical method for determining appropriate item screening criteria using cost-benefit analysis. The rationale behind cost-benefit analysis is simple: "do A if its benefits exceed its costs, and not otherwise" (Layard & Glaister, 2012, p. 1). What is not so straightforward is how we measure "cost" and "benefit." However, as Layard and Glaister (2012) put it, "The only basic principle is that we should be willing to assign numerical values to costs and benefits, and arrive at decisions by adding them up and accepting those projects whose benefits exceed their costs" (p. 2). In economics, where cost-benefit analysis originated, there are typically many variables at play, and the analysis can become fairly complex. In determining item screening criteria, however, the task boils down to weighing the cost associated with screening out items (by applying a given set of screening rules) and the benefit of measurement precision that would result from the items that remain in the pool. Here by "cost" we mean screening out items that have already been developed, not the cost of item development, although in a sense the literal cost associated with item development is subsumed in the item screening process. Specifically, by examining the effects of various combinations of item difficulty and discrimination screening criteria on a desired outcome (e.g., pass/fail classification accuracy), one can arrive at an appropriate solution given the examinee population, item characteristics, and purpose of the exam. This approach is intended to help practitioners empirically determine which screening criteria are best for their testing program, eliminating the need to rely on general guidelines. In the following sections, we describe the process for determining optimal screening criteria and illustrate the method in the context of a high-stakes medical certification exam.

Analytical Framework

Determining appropriate screening rules is ultimately a matter of weighing the potential costs and benefits of any particular set of screening criteria. As screening criteria become more stringent, more items will be removed from the live pool (cost). At the same time, the quality of the remaining items in the pool will improve. This improved item quality ultimately leads to more accurate scores and classification decisions (benefit). The challenge is in determining which set of screening criteria achieve an optimal balance of cost and benefit.

Cost-Benefit Simulation

Simulation is an effective tool to empirically connect each potential set of screening criteria to its associated costs and benefits. To ensure that the simulation reflects the results that are likely to be observed during an operational exam administration, it is important to gather the IRT parameters from a large number of pretest items. The use of IRT is not a requirement for cost-benefit analysis, but it makes the simulation and analysis easier. It is critical that item statistics come from pretest items, because the live item pool will only include items that meet the current set of screening criteria, and it will therefore be impossible to explore the impact of less stringent criteria. Once the potential item pool is collected, classical difficulty (p) and discrimination (e.g., biserial correlation, r_{bis}) values must be calculated for all items in the pool. For items calibrated in a 2- or 3-parameter IRT model, it may be reasonable to calculate "true" classical statistics based on the IRT parameters and actual examinee ability distribution. These values can be used to simulate responses from a large number of examinees for each item. Once these response strings are simulated, one can compute true classical statistics for each item. This approach helps to eliminate noise in the classical statistics that will be observed using pretest responses alone, which are often based on small samples. With the pretest item parameters, classical statistics, and examinee ability distributions available, it is now possible to connect each set of screening criteria to a specific cost and benefit.

Choosing Relevant Screening Criteria

A key component of cost-benefit analysis is specifying what item screening criteria combinations should be examined in the simulation. In principle, one could investigate all p-values between 0.0 and 1.0 and all discriminations between -1.0 and 1.0. As a practical matter, the recommendations found in the literature may be helpful in simplifying the analysis. For most exams, it will probably be appropriate to examine biserial discriminations between 0.0 and 0.30. For difficulty, p-values less than 0.30 and greater than 0.80 may warrant examination. Ultimately, the selection of appropriate ranges is at the discretion of each practitioner. However, it may be practical to err on the side of including too many potential screening rules, as there is little additional work associated with examining additional rules.

Building a Theoretical Item Pool for Each Set of Screening Criteria

Once the relevant screening criteria have been defined, one can apply all combinations of difficulty and discrimination screening rules to the pool of items. Any item that does not meet the specified criteria should be removed from the potential item pool. This process yields one theoretical item pool for each combination of potential screening rules. In essence, this step simulates the composition of the live item pool one would observe if they implemented different screening rules.

Sampling Items to Build Forms, Simulate Responses, and Estimate Parameters

Once these theoretical item pools are built, one can sample items from each to build simulated test forms based on that specific combination of screening rules. It is important to note that these simulated test forms should contain the same number of live (i.e., not pretest) and anchor items as the operational test form of interest. The designation of anchor items allows examinee abilities and item parameters to be linked back to the operational scale. Next, using the person ability distribution, one would simulate examinee responses to each test form. Finally, one would calibrate those responses using the operational IRT model to estimate item parameters and examinee abilities. Again, these estimates reflect what one would observe using different combinations of screening rules. The specific number of simulated test forms is at the discretion of each practitioner, but we have observed stable results using as few as 100 test forms for each set of screening rules.

Computing Cost and Benefit for Each Set of Screening Criteria

Having created the theoretical item pool for each set of screening criteria and calibrated the simulated test forms, one can finally calculate the optimal cost and benefit associated with each set of screening criteria. The cost is simply the proportion of the original item pool that is removed when the screening criteria are applied. So, for example, if 95% of the items in the item pool meet or exceed the screening criteria, those criteria "cost" 5% of the pool. The benefit is somewhat more complicated. If the purpose of the exam is to place examinees appropriately on the ability scale, one must calculate the difference between the true and observed ability estimate for each simulated examinee. The mean absolute difference between examinees' true and observed abilities becomes the measure of the benefit of the screening criteria. Alternatively, if the purpose of the exam is to classify examinees into discrete groups, it may be more appropriate to compute classification accuracy. Specifically, one can apply the operational cut score to the true (i.e., generating) and observed (i.e., estimated) abilities to compute the proportion of matching classification decisions. In this case, the percent of examinees correctly classified is the "benefit" of a particular set of screening criteria. After the costs and benefits are calculated for each set of screening criteria, one will be able to examine these tradeoffs and select an appropriate set of screening criteria for their particular exam program.

Selecting Appropriate Screening Criteria

After the simulation is complete, and the cost and benefit of each potential set of screening rules have been determined, practitioners must select the specific set of screening rules that makes the most sense for their exam. Although different exams may have very different requirements, screening criteria will typically be selected using a cost constraint, a benefit constraint, or an optimal balance of the two.

Cost Constraint

A cost constraint exists when practitioners are seeking the greatest benefit with a maximum fixed cost. This will frequently be the case when a testing program is limited in the number of items that it can easily produce each year or when those items are very expensive to develop. In this case, it may be reasonable to specify a cost constraint such that no more than, say, 20% of the items could be removed from the pool.

In Figure 1, we see a fictitious example of the relationship between the percent of items removed after screening (cost) on the x axis and the percent correct classification (benefit) on the y axis for 20 potential sets of screening criteria (circles). Under the 20% cost constraint, practitioners should select the gray set of screening criteria to the left of the dashed line, as it provides the greatest benefit (percent correct classification) of all options with fewer than 20% of items removed.

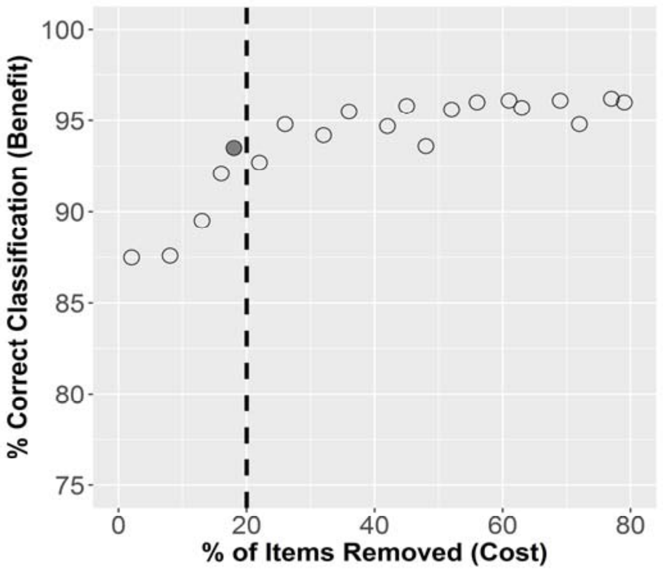

Figure 1. Selecting screening criteria using a cost constraint

Benefit Constraint

A benefit constraint exists when some minimum benefit must be achieved regardless of cost. This might occur when the consequences of misclassification are extremely high, as is the case for many licensure and certification exams. In this case, practitioners may specify a minimum benefit—say, 95% of examinees must be correctly classified—and then seek out the lowest cost that achieves this goal.

In Figure 2, we see the same data as before, but now with a 95% benefit constraint. Under this constraint, practitioners should select the gray set of screening criteria above the dashed line, as it results in more than 95% of examinees being correctly classified with the fewest items removed from the pool.

Optimal Balance

Identifying the optimal balance between cost and benefit will be appropriate when testing programs have no predetermined cost or benefit constraints. This will be the case for most testing programs with a reasonably robust item pool and no absolute requirements for classification accuracy. In these cases, the nonlinear relationship between cost and benefit makes it possible to identify the point at which additional cost will yield only trivial returns in benefit. One can employ piecewise or segmented regression (i.e., regression with unknown breakpoints) to identify the breakpoint where the

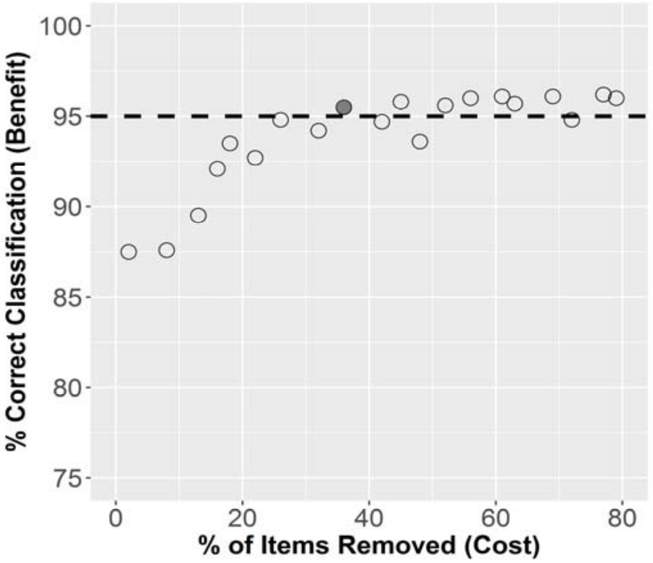

Figure 2. Selecting screening criteria using a benefit constraint

relationship between cost and benefit changes abruptly. Screening rules that produce costs and benefits around this breakpoint have an optimal balance of cost and benefit.

In Figure 3, we have fitted a segmented regression to the data. Based on these results we can see that the gray set of screening criteria at the breakpoint represents an optimal balance between cost and benefit. Applying more stringent screening rules introduces a significant loss in items with only a slight improvement in classification accuracy. Conversely, applying less stringent rules results in a significant reduction in classification accuracy with relatively small reduction in the percent of lost items. These three approaches illustrated in Figures 1-3 make it possible for practitioners to identify the appropriate screening criteria given the constraints specific to their testing program.

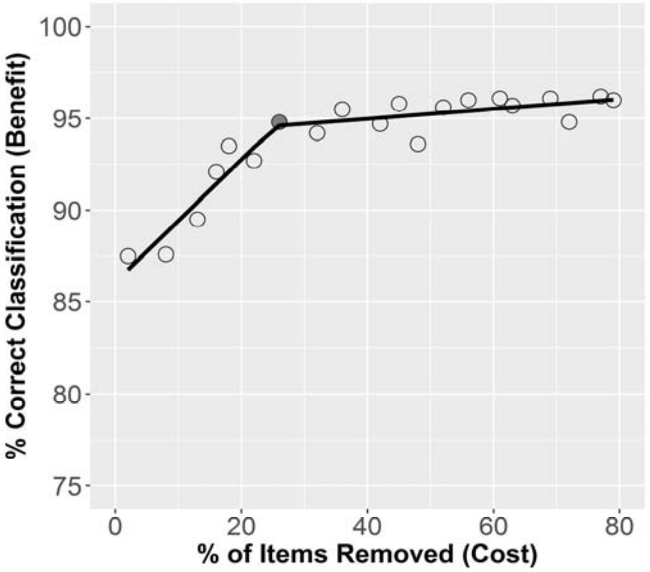

Figure 3. Selecting screening criteria using the optimal balance of cost and benefit

Real Data Example

To illustrate the value of cost-benefit analysis for identifying screening criteria, we analyzed real data from a high-stakes medical certification exam to determine the optimal screening criteria for that exam. This exam included an item pool with more than 3,000 items all calibrated with the 2-parameter logistic IRT model. We manipulated five levels of discrimination (r_{bis} ≥0.10, ≥0.15, ≥0.20, ≥0.25, ≥0.30) and four levels of difficulty (p ≤0.95, ≤0.90, ≤0.85, ≤0.80) for a total of 20 screening criteria combinations. We chose these specific values because they provide a range that spans most recommendations from the literature and agree with our earlier simulation work that suggested *our exams* could be improved by increasing discrimination (i.e., higher r_{bis}) and difficulty (i.e., lower p-values). Once we determined the item pool, examinee population, and screening criteria, the next step in the cost-benefit analysis was to conduct the simulation study.

We began the simulation by applying the 20 sets of screening criteria to the complete item pool to create 20 theoretical item pools. The largest of these pools had over 3,000 items and the smallest had just over 650 items. Then we randomly selected 200-item forms from each of these live pools. Next, using the items' operational IRT statistics as generating values, we simulated item responses for roughly 2,000 thetas with mean = -0.06, *SD* = 0.94. We calibrated these responses and linked them back to the operational scale using 60 anchor items. Both the total number of items and the anchor size mirrored our operational practices. Finally, we applied the operational passing score to the generating abilities and the equated ability estimates to produce true and estimated pass/fail decisions. Doing so allowed us to compute classification accuracy in each condition. We replicated this process 100 times for each set of screening criteria to ensure that we obtained stable results. The simulation culminated in an estimated cost and benefit for each of the potential screening rules.

Results

As expected, there was a positive relationship between cost and benefit, with more stringent screening criteria (and a larger percent of items removed after screening) resulting in higher percent correct classification. That is, selecting higher quality items from the pool was associated with greater classification accuracy. Importantly, the relationship between cost and benefit was nonlinear: initial increases in the stringency of screening criteria (and increases in cost) were associated with substantial increases in classification accuracy up to a point (around 22% of items removed from the original pool). After this point, applying more stringent criteria and removing a larger proportion of items from the pool was associated with only modest increases in classification accuracy. The cost-benefit analysis example presented here had no predetermined constraints for either cost (i.e., percent of items to be screened out of the pool) or benefit (e.g., percent correct classification). In other words, our goal was to find the optimal screening criteria solution that would maximize pass/fail classification accuracy and minimize the amount of items we would remove from the pool. Therefore, we performed segmented regression using the R package “segmented” (Muggeo, 2003; 2008) to identify and plot the breakpoint in the cost-benefit relationship (see Figure 4). Doing so allowed us to identify the turning point after which we observe no substantial added benefit at a reasonable cost.

There were two item screening criteria combinations (circles shaded in gray) near the breakpoint shown in Figure 4: a) discrimination ≥0.15, difficulty ≤0.90, which would result in 22% of items removed and 93% correct classification; and b) discrimination ≥0.20, difficulty ≤0.95, which would result in 23% of items removed and 94% correct classification. From a statistical perspective, we could champion either condition as an appropriate screening criteria solution. However, from an exam content perspective, our subject matter experts expressed interest in retaining easier items that essentially every examinee should answer correctly. Thus, we collectively agreed to discrimination ≥0.20, difficulty ≤0.95 as the appropriate screening criteria in this testing program.

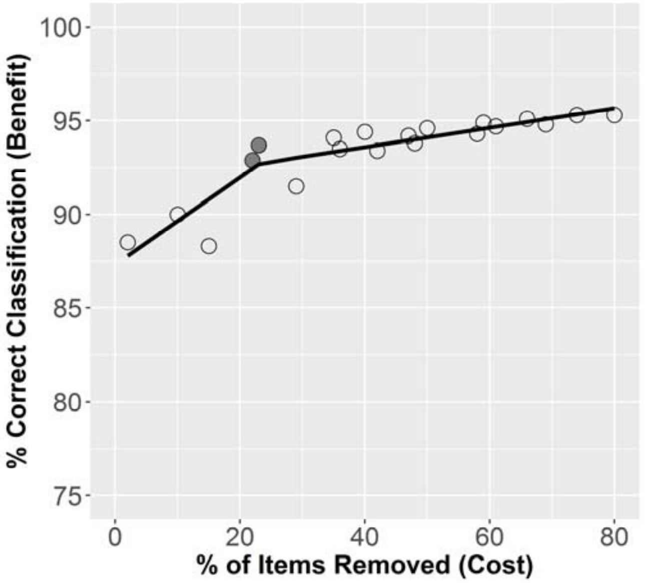

Figure 4. Cost-benefit analysis via segmented regression with one breakpoint

Discussion

The selection of appropriate screening rules can have a significant impact on the validity of inferences based on examinee scores. Unfortunately, despite the demonstrable importance of this decision, practitioners are provided with relatively little guidance on what screening criteria to adopt. As we discussed, the literature on this topic provides varied and sometimes contradictory recommendations. Moreover, such recommendations are not accompanied by citations, so they may or may not be translatable to exams with different item pools, examinee populations, and purposes. Beyond these recommendations, the literature provides limited guidance as to how a psychometrician might evaluate and ultimately select the screening criteria that are most appropriate for a given exam.

This paper presents a method for evaluating the costs and benefits of possible screening criteria including a framework for selecting the appropriate criteria given the needs of the exam. We believe that framing the selection of screening criteria as a balance between cost and benefit formalizes the tradeoffs that must be made, and provides empirical evidence in support of the ultimate selection. Our hope is that this paper has shed some light on the consequences of applying various combinations of item screening rules on an outcome of interest and will serve as a guide to assessment professionals in their own endeavors to determine appropriate item screening criteria.

To demonstrate the efficacy of this approach, we applied this method to data from a high-stakes medical certification exam. A cost-benefit analysis revealed that the optimal screening criteria were either discrimination ≥0.15, difficulty ≤0.90 or discrimination ≥0.20, difficulty ≤0.95. Both combinations of screening criteria resulted in essentially the same improvement in pass/fail classification accuracy for the least number of screened out items. As shown in Figure 4, practitioners could be too stringent if they adopted guidelines greater than the ones discussed above, as such guidelines add little improvement in the outcome for a noticeably large loss of items. Alternatively, adopting a less stringent set of screening criteria could be too lenient a choice, as the few retained items would cost the practitioner a great reduction in accuracy. This is exactly why a disciplined approach to the selection of screening criteria is so important. Ultimately, the optimal set of screening criteria will depend on the item bank depth/breadth, the blueprint, and cost-benefit tradeoffs.

Limitations

The most significant limitation of this method lies in its rather significant data requirements. For testing programs to implement this method, they require a large item pool with both pretest and live items. This will not be feasible for all testing programs. Furthermore, since the simulation relies on the fact that classical difficulty and discrimination would manifest in the items' IRT parameters, this method may be more appropriate for item pools calibrated with the 2- or 3-parameter IRT model. That said, exams calibrated using the Rasch or 1-parameter model would still be able to identify an optimal difficulty screening rule using this approach. Despite the data requirements, we believe that the use of a data-driven method for selecting screening criteria is preferred whenever possible.

Another limitation of the cost-benefit analysis method described and demonstrated here is that items were randomly selected to form 200-item test forms regardless of content. Ideally, the selection of items should follow the operational exam blueprint so as to preserve not only the structure of the exam, but also the distribution of items among content areas. The downside to this approach, however, is that the simulation would become much more complex, and this otherwise straightforward method may lose some of its appeal to practitioners. To remedy this limitation, we examined the distribution of item parameters across all content areas outlined in the blueprint and found no practically significant differences among content areas in terms of item quality. That is, if we were to apply a more stringent set of screening criteria, no content area would be disadvantaged.

Conclusion

The present study provides an empirical method of determining what item screening criteria to use in practice. The method is straightforward to implement and provides empirical evidence to support the ultimate selection of screening criteria. We hope the illustration presented here serves as a good example of the rationale, logistics, and, most importantly, the value of using cost-benefit analysis to determine appropriate item screening criteria in a variety of assessment contexts.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Assessment Systems Corporation. (2017). User Manual for Iteman 4.4. Minneapolis, MN: Author.

Chubbuck, K., Curley, W. E., & King, T. C. (2016). Who's on first? Gender differences in performance on SAT® Critical Reading items with sports and science content (ETS Research Report No. RR-16-26). Princeton, NJ: *Educational Testing Service.*

Clauser, J. C., & Hambleton, R. K. (2017). Item analysis for classroom assessments in higher education. In C. Secolsky & D. B. Denison (Eds.), Handbook on measurement, assessment, and evaluation in higher education (2nd ed., pp. 355-369). New York, NY: Routledge.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart, & Winston.

Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, NJ: Prentice-Hall.

Ilic, D., Bin Nurdin, R., Glasziou, P., Tilson, J. K., & Villanueva, E. (2014). Development and validation of the ACE tool: Assessing medical trainees' competency in evidence based medicine. *BMC Medical Education*, 14, 1-6.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30, 17-24.

Layard, R., & Glaister, S. (2012). Cost-benefit analysis (2nd ed.). Cambridge, UK: Cambridge University Press.

Liu, J., Harris, D. J., & Schmidt, A. (2007). Statistical procedures used in college admissions testing. In C. R. Rao, & S. Sinharay (Eds.), Handbook of Statistics, Volume 26: *Psychometrics* (pp. 1057-1094).

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22, 3055-3071.

Muggeo, V. M. R. (2008). segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News, 8/1, 20-25. Retrieved from https://cran.r-project.org/doc/Rnews/

Nunnally, J. C. (1967). Psychometric Theory. New York, NY: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric Theory (3ed ed.). New York, NY: McGraw-Hill.

Terrell, C. D. (1982). Table for converting the point biserial to the biserial. *Educational and Psychological Measurement*, 42, 983-986.

The College Board (2016). SAT Technical Manual. Retrieved from http://www.sde.ct.gov/sde/lib/sde/pdf/evalresearch/sattechmanualcompletefinaltextforstates.pdf

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4rd ed., pp. 307-354). Westport, CT: Praeger.

Varma, S. (2010). Preliminary item statistics using point-biserial correlation and p-values. Retrieved from https://eddata.com/wp-content/uploads/2015/11/EDS_Point_Biserial.pdf

Youngman, M. B. (1979). A comparison of item-total point biserial correlation, Rasch and alpha-beater item analysis procedures. *Educational Studies*, 5, 265-273.

Zumbo, B., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), The SAGE Handbook of Quantitative Methodology for the Social Sciences (pp. 73-92). Thousand Oaks, CA: Sage Press.

Citation:

Bashkov, Bozhidar M., Clauser, Jerome C. (2019). Determining Item Screening Criteria Using Cost-Benefit Analysis. *Practical Assessment, Research & Evaluation*, 24(2). Available online: http://pareonline.net/getvn.asp?v=24&n=2

Corresponding Author

Bozhidar M. Bashkov, Measurement Scientist
American Board of Internal Medicine
510 Walnut Street, Suite 1700
Philadelphia, PA 19106

email: bbashkov [at] abim.org