Item Quality Analytics 101: Dipping a Toe in the Water

ITCC Blog Post July 2023. Written by Elwood Fischer, member of ITCC

Introduction

To be honest, it took a while for our program to rediscover item analytics after some major turnover. After our new team learned basic principles of writing quality multiple-choice items, we wanted to see how our existing items measured up. Unfortunately, our small program didn't have resources for actual psychometric help. But we were delighted to discover that a couple of measures can go a long way, especially with help from colleagues in a knowledgeable community like the ITCC. In this post, I'm sharing what I learned in those early forays, with the goal of demystifying item analytics for people new to certification, and inspiring you to step into doing them. The numbers are more accessible and can tell you a lot more than you might think – and that's actually quite fun!

I am not a psychometrician, and careful psychometrics are important, especially for high-stakes exams. For your safety, I've included input from some of those previously-mentioned colleagues, to help you know what to guard against and when to get professional help. All right, then, let's dip a toe in the water!

Item Analytics are calculations done to evaluate the quality of individual items within an assessment. They may be used as part of an overall evaluation of exam quality, but each calculation focuses on a specific item.

Using item analytics may seem daunting, especially for small programs that don't have many specialized roles on their team, but it doesn't have to be. Many item-banking and reporting tools include basic item analytics. Even Excel has formulas for them. Understanding how to apply some basic item analytics can greatly improve your exams and your program.

This blog post will discuss two calculations that I found helpful and powerful as a non-psychometrician working to improve my certification program. Used in concert, p-value and point-biserial correlation (PBC) go a long way towards identifying exam items that contribute to your exam achieving its goals of testing the knowledge, skills, and abilities (KSAs) at the right level for your target candidates, and of distinguishing between qualified and unqualified candidates.

P-value

P-value, also called item difficulty index, is the average score on the item of respondents to the item. For dichotomous items, meaning items marked as either correct or incorrect, this will be the proportion of respondents to an item who answered it correctly. Items that are part of an adaptive testing chain, or whose grading allows partial credit, would require more sophisticated calculations.

P-values for dichotomous items will range from 0 to 1, meaning from nobody who answered the question got it right to everybody who answered it got it right. The target difficulty is program specific. For example, my program used 0.3 < p-value < 0.9, while Microsoft uses 0.2 < p-value < 0.9. They retain more difficult questions, reflecting the higher rigor of their program and processes. Items with p-values outside your target range may warrant double-checking the item quality. A question may have a low p-value because your candidates find it confusing, and an item with a high p-value may be poorly written in a way that makes the correct answer obvious. Sending the questions for SME review will provide feedback to help you decide whether to revise, retain as-is, or remove items, consistent with the goals of your program and exam.

Once an exam is released, you can support its maintenance by periodically comparing the p-value of an item in a recent time period with the p-value of that item in previous periods. If the p-value increases significantly over time, it's possible that the measured KSA has become more mainstream since the item was written. Or possibly the item has been overexposed because the question has leaked into publicly available material, such as training or brain dump websites, or simply because the item has appeared on many exams over time. On the other hand, if the p-value drops, you may want to check whether the KSA being tested has changed. For example, if the way a feature works has changed without updating items which use this feature, candidates trained on the new product will find it harder to answer those questions correctly.

The p-value can help verify that the difficulty of an exam item contributes to the intended difficulty of the exam for your target audience. But it has a significant limitation; it can't indicate whether the "right" candidates are getting the question correct. The next metric we discuss helps address this issue.

Point-Biserial Correlation (PBC)

An excellent complement to p-value is an item's point-biserial correlation (PBC). PBC is the correlation between the candidate's item score and their overall score on the exam.

An underlying data assumption is that the PBC can be applied to data when one variable is dichotomous (*e.g.* right/wrong, with one correct answer) and the other is continuous (*e.g.* test scores). Under these conditions, the point-biserial correlation is similar to the Pearson product moment correlation.

Correlation coefficients have two parts: the value and the sign. The value relates to the strength of the relationship — the closer to an absolute value of 1, the more related the two values are; as one increases or decreases, so does the other. The closer to 0, the less related the values are; as one increases, the other could increase, decrease, or stay the same. The sign provides information about the direction of the relationship. If positive, they are moving in the same direction, if negative, they are moving in opposite directions.

So a PBC value of +1 means respondents scored well on this exam in exact proportion to their overall score on the exam. A PBC of zero doesn't discriminate at all between qualified and unqualified candidates. And a PBC of -1 means candidates got the question correct in exactly inverse proportion to how well they scored on the exam (more about that later).

Higher point-biserial correlations are better, and the higher-stakes the exam, the higher you'll want your point-biserial correlations to be. Typically, a PBC higher than 0.2 indicates a "good" item, and below 0.1 is problematic. Items with low PBCs do not distinguish between qualified and unqualified candidates.

So what does a negative PBC tell you? Most commonly, items with a negative PBC have been mis-keyed, so higher-scoring respondents are answering the question correctly but having the item scored as incorrect. If the negative-PBC item is keyed correctly, it's possible the question does not rule out edge cases that make one of the distractors true, and high-performing candidates are selecting that edge case.

Running a PBC on each response can be very helpful to improve question quality. Distractors should have significant and negative PBCs, as well-qualified candidates should NOT select them. A distractor with a strong positive PBC will significantly reduce the PBC of the item. And a distractor with a PBC that is near zero is not contributing to the discriminatory power of the item. Consider editing the distractor, or even dropping it. Recent research has shown that questions with 2 distractors tend to perform as well as those with 3.

Higher-level applications

These two statistics, when used together, can provide an easy and powerful entry-point into item analysis. It's obvious that using them can greatly improve the quality of your items. And you can do so much more. Here are a few possibilities I think you might find interesting:

- Item analytics can be an important input to other processes, such as ensuring the equivalency of different exam forms or aligning a form with the exam's scoring scale.
- If you run a pilot on the items before finalizing your forms, the item analytics can help you avoid using poor-quality items. You can fix the item before using it, or decide not to use the item. Graphing items using p-value on one axis and PBC on the other can help quickly identify potentially problematic questions and can help guide and prioritize a response. Including the target window for each dimension can help provide an indication and visual display of the overall quality of an exam as reflected by these two measures.
- You can also examine the poor-quality questions as a group, and see if you can identify what they have in common. For example, topics whose items have lower analytics than others may need improved preparation material and messaging. Using different colors for each topic on the p-value vs PBC plot will make topic-level issues obvious. Another issue may be questions of the same question type or structure performing poorly. In this case you might either revise the questions to another type or provide training or preparation opportunities so candidates are comfortable with questions of that type.
- If you are running item analytics for localized versions of the same exam, you might compare the analytics for the same item in different languages (if you have enough data for a language). This may provide an easy first pass in identifying poorly-localized items or distractors, or items that are culturally inappropriate. But be aware there are psychometrically stronger methods for doing this.
- Finally, if your item analytics bring to the surface issues that will be deferred to the next revision or major revision of the exam, be sure to document the information along with

customer feedback, product updates, and other information that will inform the content and timing of the revision.

Troubleshooting

When an exam has a high number of questions with low p-values and/or PBCs, I've found the following may be helpful:

- I start by double-checking the quality of the data.* Candidates with widely-varying experience and expertise will weaken correlations. Also, the relevance of item analytics can be undermined if your data includes candidates that do not reflect the exam's intended audience. For example,I learned that student data on a professional exam is notoriously unreliable. Running the analysis on those candidates who fit your audience profile will yield more accurate statistics.
- Next, I might look for patterns in item writing to see if the item writers have fallen into a pattern or a small set of patterns for questions that are not working for your intended candidates.
- If the data and item quality check out, I might take a look at my exam blueprint. The point-biserial correlation assumes the exam as a whole represents a single, cohesive body of knowledge. If this is not the case, then correlation between correct answers in one topic (or a set of topics) will not necessarily be correlated with getting correct answers in the non-related topics. A low average PBC could result. In such cases, if the topics are confirmed to be important to include, I might consider adjusting training and messaging around the low-performing topics.

*Confession: In my first attempt at using PBC, I actually started by asking a more experienced industry colleague, and their first question was whether we had included student data. A quick, early check with a psychometrician can save a lot of time and help you know whether you can proceed with confidence or need to adjust your approach.

Conclusion

P-values and point-biserial correlations of items can provide a good start towards improving exam quality. They can help you proactively identify and correct issues in an exam, rather than simply reacting to or defending against customer complaints after exam release. And after you and your program are comfortable applying these, there's so much more!

Caveats

This blog post, while perhaps helpful to demystify item analytics for a non-technical audience, oversimplifies at several points which will be worth examining later, including:

- Please note that these two calculations focus narrowly on the difficulty and the distinguishing power of the item. They do not measure item validity, for example, which would require correlating performance of the item with an external measure of validity.

- Assumptions of the point-biserial correlation include that the respondents' scores on an item are normally distributed. It's good practice to verify this using a scatterplot (number of respondents vs score). You can also check this by calculating the Spearman's correlation; for normally distributed data the two values should be similar.
- Finally, the p-value and point-biserial correlation, and this discussion of item analytics, is grounded in classical test theory and therefore has all the limitations of that approach which item response theory helps address. If your program can use item response theory, this post's basic information about p-value and PBC will hold, but you'll have better tools available for the "higher-level applications" and for troubleshooting than what is discussed here.