How to Write a Multiple-Choice Question

(According to the research.)

Introduction

This paper began with me asking the question, "How can research help me write items for SMI?" (Should this paper ever reach the eyes of the unfamiliar, SMI is a test that strives to measure a student's math ability.) I understand that science rarely has the perfect solution for problems in writing or education. Expert judgments, product-specific standards, and personal anecdotes all have their place.

Nevertheless, I encountered dozens of researchers who have studied the science of multiple-choice questions. Most notably, Thomas Haladyna out of Arizona State University West is the most prominent current researcher. I draw liberally from his work, which is by far the most cited in the literature.

Number of Choices

I'll begin with the work to determine how many choices a multiple-choice question should have. In other words, do we really need an A, B, C, and D? I found it to be one of the few counterintuitive results. There have been many studies trying to find the optimal number, and current research suggests that two is fine, three is optimal, and four is, believe it or not, excessive (Landrum, Cashin, et al., 1993).

It turns out all you need is the right answer and one or two plausible wrong answers to satisfactorily discriminate those who know the content from those who don't. Sure, a student could make a mistake other than the distractors provided, but when the student commands the material down to a coin flip, that's actually a statistically perfect item.[1] It's no wonder that despite most high-stakes assessment items having four options, typically only one or two of the distractors are actually plausible (Haladyna & Downing, 1989).

One of the reasons shorter items are better is because more of them can fit on a page. That not only saves space, that saves time. The items take writers less time to write and students less time to solve (Haladyna & Downing, 1993). Note that it's impossible to separate the item writing process from the items themselves.

[1] I'm referring to the Rasch model, the statistical method SMI (and many other tests) use to determine a test's validity. How to measure a test's validity, along with the strengths and weaknesses of the Rasch model, exceed the scope of this paper.

Having more items on a test also makes the scores more reliable and lets the test cover more of the content domain (Ebel, 1981).

On a technical note, having fewer options does increase the baseline score of random guessing from 25% to 33% (if there are three options) or 50% (if there are two). However, given enough questions, this does not affect a test's discriminatory ability.[2] Fewer options especially helps measure high-achieving students who have become skilled test takers without always knowing the content (Williams & Ebel, 1957).

Homogenous Choices

Initially, I assumed that if only one good distractor was enough, then the composition of the other distractors wouldn't matter. My instinct failed me. Test takers use clues about the choices when they don't know the answer. ("B is longer than the other choices, so I bet that's the answer.") This can affect both an item's difficulty and its reliability. There really is a "go with the longest answer" heuristic that students might follow, for instance. Making all of the choices the same length averts this bias and improves the validity of the test's measurement (Towns, 2014).

Our brains are in fact full of biases. If one option is unusually technical, worded strangely, or in general looks different, test takers pick up on it. The test item can therefore be solved by either knowing the content domain or having a keen test taking ability (Haladyna, 2004). In one study out of the Netherlands, researchers gave written driving tests to hundreds of high school students. Some students saw a version of each question with homogeneous options whereas others didn't. Sure enough, the heterogeneously-written version of each question tested a little bit easier (Ascalon, Meyers, et al., 2007).

The order of the distractors can matter, too. In another study, researchers gave high schoolers an ACT-like math test. Half the students saw the questions normally, and the other half saw the distractors in a random order. For most students, the difference was irrelevant. Mixing up the options didn't make it any harder to find the right one. However, this wasn't true for lower-ability test takers. Randomly-ordered distractors made a question more challenging for them. (Huntley & Welch, 1993).

Simple Language

So far we suspect that multiple-choice questions measure knowledge more accurately when the choices have the same length and structure and are listed in a

[2] This result is mathematical, and I'll leave it as an exercise to the reader.

logical order. Common wisdom is also to write the questions and choices as simply as possible. The AERA's *Standards for Educational and Psychological Testing* asserts that simple language reduces the influence of reading ability when that's not what the test is supposed to assess (AERA, et al., 1999). It's an assumption we follow at HMH: keep the language simple.

The notion seems right, for sure. But does it hold up to scrutiny? Jamal Abedi out of UCLA conducted a study where his team gave a math test to middle school students. Half the students received the normal test, and the other half received one written in simplified English (Abedi, Lord, et al., 2000):

- Nonmath terms that a student might not know were changed (e.g. "census" became "video game").
- Passive voice was changed to active voice. (I'll just accept the irony that my sentence came out in passive voice.)[3]
- Long nominals were shortened (e.g. "last year's class president" became "president").
- Conditionals were replaced with separate sentences. ("If Johnny delivers *x* newspapers…" became "Johnny delivered *x* newspapers.")
- Relative clauses were removed. ("A report that contains 64 sheets of paper" became "He needs 64 sheets of paper for a report.")
- Complex question phrases were replaced with simple question words. ("At which of the following times…" became "When….")
- Impersonal presentations were made concrete. ("The weights of 3 objects were compared" became "Sandra compared 3 objects.")

For non-ELL students, the difference in test scores was negligible. If a student's first language is English, changing the complexity of a question's wording doesn't make too big a difference. However, the same is not true for ELL test takers. Simplifying the English increased their average test score from 34.5% to 36%, a statistically significant jump relative to the sample size. In other words, simplifying the language narrowed the gap between ELL and non-ELL test takers.

Positively-Worded Questions

Another often-cited guideline is to write questions as "which is true?" as opposed to "which is false?" (Haladyna & Downing, 1989). In 1993, Israeli researcher Pinchas Tamir challenged this claim. He gave a high school biology test to students in both Australia and Israel. The Israelis' test was translated into Hebrew and checked independently. Some students saw the positively-worded version of each question and others saw the negatively-worded one. The difference was stark. If

[3] There was one notable study on humor in testing. In most cases humor doesn't help, but it probably doesn't hurt either (McNorris, F., Boothroyd, R., et al., 1997).

the question was easy, there was no difference in performance between the two versions. However, if the question had a high cognitive demand, the guideline proved correct. Students were more likely to get the negatively-worded version wrong (Tamir, 1993).

The research is not in full agreement, however. Three years later, two researchers analyzed hundreds of health professional exam results. They compared the performance of questions worded positively against those worded negatively. Statistically, there was no difference between the two (Rachor & Gray, 1996). Incidentally, I've noticed that using tests from the medical field is a trend in the research. My guess is it's because the industry is replete with regulations and testing and the stakes can be unusually high.

We are likely doing the right thing by avoiding negatively-phrased questions. Our audience is closer to the high school biology student than the medical professional. The research is clear, though, that a negatively-worded question occasionally makes more sense. It's worth remembering that if a question is simple, phrasing it negatively is unlikely to make much of a difference.

"None of the Above" and "All of the Above"

The question of when and whether to use "none of the above" is an especially active area of multiple-choice test research. Perhaps unsurprisingly, the research is mixed. Some studies found that "none of the above" increases an item's difficulty (e.g. Forsyth & Spratt, 1980); others didn't (e.g. Rich & Johanson, 1990). Some studies found that "none of the above" increases an item's discriminatory ability (e.g. Oosterhof & Coats, 1984); others didn't (e.g. Frary, 1995)

It appears that when "none of the above" is an option, test takers are compelled to actually solve the problem rather than use non-content skills (Dochy, Moekerke, et al., 2001). However, there's a problem when it's the correct answer. Canadian researcher David DiBattista performed an experiment where he gave college students a general knowledge test. Students received one of five versions of the test, all having "none of the above" in different places, sometimes being the correct answer and other times not.

DiBattista compares the answer choices to a police lineup.[4] On average, a witness will spot the perpetrator in a lineup if he or she is actually in it. However, if the perpetrator is absent (i.e., "none of the above"), witnesses select whoever looks the closest and think they picked the right person. Students, even ones who

[4] I don't know how many research papers on multiple-choice testing I read, but this was the only one where it got weird.

actually know the content, are sometimes unable to deduce the right answer if it isn’t one of the choices (DiBattista, Sinnige-Egger, et al., 2013).

Unlike “none of the above,” which poses problems but can be employed carefully, “all of the above” seems universally problematic. Having “all of the above” as an option naturally cues test takers that it’s the right answer (Harasym, Leong, et al., 1998). It makes for a lousy correct answer and a misleading distractor, overall harming the reliability of a test that contains it. Test writers are wise to avoid “all of the above.”

Conclusion

It’s impossible for guidelines to be perfect. A clever item writer could think of questions where plausible distractors do not look homogenous or “all of the above” makes perfect sense. But I’ll finish with a study where researchers examined all of the test items one medical school class received throughout the year. They reviewed each item for whether it followed the rules I’ve described (and many I didn’t) for writing a multiple-choice question, and compared them against the student’s scores. When an item had errors, it hurt the student’s performance “without providing additional discrimination between higher- and lower-performing individuals” (Pate & Caldwell, 2014).

There are many details I did not get into. There’s research on when multiple-choice is better than constructed response, what type of multiple-choice question is best, how to train item writers, whether it matters if a stem has fill-in-the-blank, which statistical test is best, and so on and so on.

If you make time for just one reference, Thomas Haladyna, along with Steven Downing out of the University of Illinois at Chicago, published a “taxonomy of multiple-choice item-writing rules” in 1989. It’s a comprehensive summary of all the research up till then on writing a multiple-choice test item (Haladyna & Downing, 1989). The article is approachable and interesting. They updated the list in 2002, and while their conclusions stay largely unchanged, they address new research and refine some arguments (Haladyna, Downing, et al., 2002). It’s impossible to write a perfect multiple-choice question, but fortunately there’s plenty of research to help us try.

References

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16-26.

AERA, APA, & NCME [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education]. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Ascalon, M., Meyers, L., Davis, B., & Smits, N. (2007). Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty. *Applied Measurement in Education*, *20*(2), 153-170.

DiBattista, D., Sinnige-Egger, J., & Fortuna, G.. (2013). The "None of the Above" Option in Multiple-Choice Testing: An Experimental Study.*The Journal of Experimental Education*, *82*(2).

Dochy, Moekerke, De Corte, and Segers. (2001). The assessment of quantitative problem-solving with "none of the above"-items. *European Journal of Psychology of Education*, *26*(2), 163-177.

Ebel, R. (1981). *Some advantages of alternate-choice test items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Forsyth, R. & Spratt, K. (1980). Measuring problem solving ability in mathematics with multiple-choice items: The effect of item format on selected item and test characteristics. *Journal of Educational Measurement*, *17*(1), 31-43.

Frary, R. (1995). The None-of-the-Above Option: An Empirical Study. *Applied Measurement in Education*, 4(2), 115-124.

Haladyna, T. (2004). Guidelines for Developing MC Items. In *Developing and Validing Multiple-Choice Test Items, Third Edition* (pp. 97-126). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. & Downing, S. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1)*,* 37-50.

Haladyna, T. & Downing, S. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement*, *53*(4), 999-1010.

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309-334.

Harasym, P., Leong, E., Violato, C., et al. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation and the Health Professions*, *21*(1)*,* 120-133.

Huntley, R. & Welch, C. (1993). *Numerical Answer Options: Logical or Random Order?* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Landrum, R., Cashin, J., & Theis, K. (1993). More Evidence in Favor of Three-Option Multiple-Choice Tests. *Educational and Psychological Measurement*, *53*, 771-778.

McNorris, R., Boothroyd, R., & Pietrangelo, D. (1997) Humor in Educational Testing: A Review and Discussion. *Applied Measurement in Education*, *10*(3), 269-297.

Oosterhof, A. & Coats, P. (1984). Comparison of difficulties and reliability of quantitative word problems in completion and multiple-choice formats. *Applied Psychological Measurement*, *8*(3), 287-294.

Pate, A. & Caldwell, D. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning*, *6*(1), 130-134.

Rachor, R. & Gray, G. (1996). *Must All Stems Be Green? A Study of Two Guidelines for Writing Multiple Choice Stems.* Paper presented at the Annual Meeting of the American Educational Research Assocation. (New York, NY, April 8-12).

Rich, C. & Johanson, G. (1990). *An item-level analysis of "None-of-the-above."* Paper presented at the annual meeting of the American Educational Research Assocation (Boston, MA, April 16-20).

Tamir, P. (1993). Positive and negative multiple-choice items: How different are they? *Studies in Educational Evaluation*, *19*, 311-325.

Towns, M. (2014). Guide to Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *Journal of Chemical Education*, *91*(9), 1426-1431

Williams, B. & Ebel, R. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabluary test items. In *The 14th Yearbook of the National Council on Measurement in Education* (63-65). Washington, DC.