THE OHIO STATE UNIVERSITY

Looming Large: How do Correct Answer Choice Lengths Impact Item Characteristics?

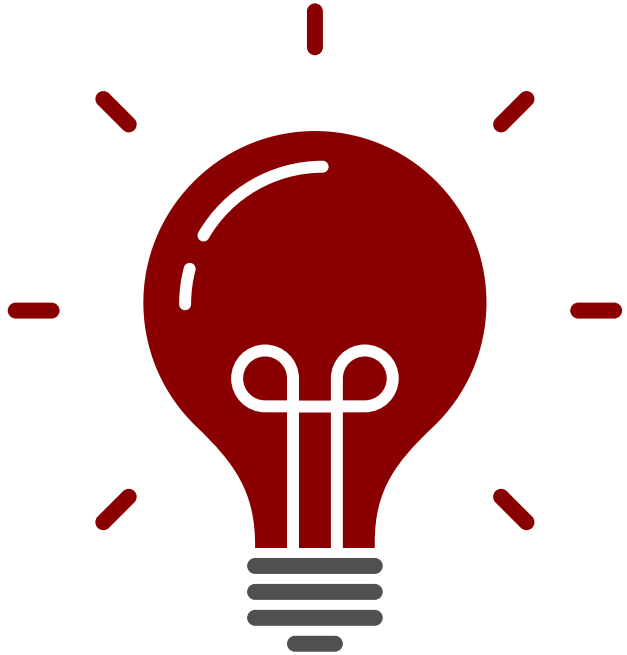
Rebecca Berenbon, Bridget McHugh, & Abena Anyidoho

Introduction

- When writing multiple-choice test questions, it is often difficult to write distractors that are both plausible and clearly incorrect.
- Guidelines for writing multiple-choice items suggest item writers should "keep the length of choices about the same" (Haladyna, 2004. p. 116)
- Item writers often struggle to write distractors that are as detailed as the correct answer.
- When items violate this guideline, correct answers may stand out to "test-wise" examinees, leading to an unfair advantage (Downing, 2002; Haladyna, 2004).
- When the correct answer is the longest option, the item is expected to be easier than intended (Rush et al., 2016).

Background

- When the correct answer choice is the longest option, the item tends to be easier (Haladyna & Downing, 1989; Pham et al., 2018)
  - Pham and colleagues observed an average item difficulty of p= .66 for flawed items (correct answer choice was the longest option) and an average item difficulty of p= .62 for unflawed items.
- Effect on discrimination is less-frequently studied and findings are inconsistent (Haladyna & Downing, 1989; Pham et al., 2018)
  - Haladyna & Downing's metaanalysis only reported 1 study that reported on item discrimination: Board & Whitney (1972) found that low-achieving students benefitted from this item writing flaw while high-achieving students did not.
  - In Pham and colleagues' later study, they found no effect on discrimination.



Background

- In previous research, items are usually classified as *flawed* (correct answer is longest) or *unflawed*
  - Could binary classification system mask more complex relationships?
- Items are compared using average difficulty of flawed vs. unflawed items (e.g., Rush et al., 2016) or matched pairs (e.g., Pham et al., 2018)
  - Previous studies have not incorporated SMEs' expectations of item difficulty

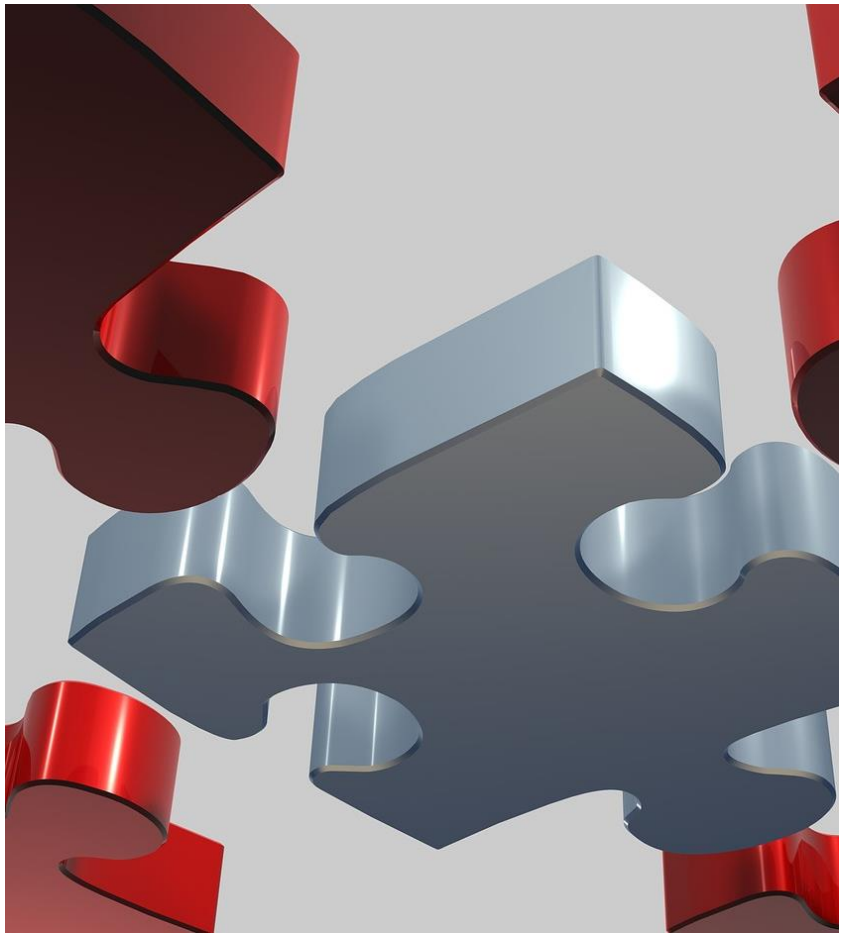
An extreme example of a flawed item:

What is the definition of the word “oligarchy”?

a) a large quantity
b) a woman’s hat
c) a wooden arrow
d) a small group of people having control of a country, organization, or institution

Research Questions

1. How often do correct answer choices **differ significantly in length** from their distractors?
2. How does the **relative length of the correct answer choice** correlate with **item properties**, including a) difficulty and b) point-biserial discrimination?



Method

- **The tests**
  - 5 courses in large-scale assessment program
  - Tests written by small groups of current secondary and post-secondary instructors, each led by a trained facilitator
  - 91-99 items on each test
  - 4 answer options for each test
  - 464 test items and 1,856 answer choices analyzed
- **Data**
  - Field testing data used to calculate **item difficulty** and **discrimination**
  - 1,209-3,042 examinees took each test (M=2,062.6, SD= 585.1)

Method

- **Analysis**
  - Calculated standardized lengths of each answer choice
    - **Length** = number of characters including spaces
    - Standardized for each item such that **average answer choice length** = 1.0
  - Using our previous example:

Answer Choice (*= correct) *What is the definition of the word "oligarchy"?*	**Length (characters)**	**Std. Length**
a) a large quantity	16	0.5
b) a woman's hat	13	0.4
c) a wooden arrow	14	0.5
d) a small group of people having control of a country, organization, or institution*	81	2.6

Method

- **Analysis**
  - Used SMEs' modified Angoff judgments (Plake et al., 2012) to quantify **expected** item difficulty
    - SMEs answered whether they believed that a borderline-proficient student would answer the item correctly (Yes/No)
    - Between 9 and 14 SMEs rated each item (M= 12.5, SD= 0.9)
    - Calculated % "Yes" per item
  - Also calculated **actual item difficulty** (i.e., p value; percent correct) and item discrimination (point-biserial correlation) for each item
  - All analyses performed in R

Method

- **Analysis**
  - RQ1
    - Examined descriptive statistics for answer choice length
    - Used ANOVA to compare lengths of correct answers vs. distractors
  - RQ2
    - Used standardized correct answer lengths to predict actual item difficulty (p) controlling for aggregated standard setting ratings
    - Used standardized correct answer lengths to predict point-biserial discrimination values for each item

Results

Table 1

Relative length of correct answer choice

	n	%
Shortest	95	20.3
Neither shortest nor longest	269	57.5
Longest	104	22.2

- Correct answer choice length was around the average in most cases
- Only shorter and longer 20% of the time
  - Probable, given four answer options
- Min = .36 (a third of the average length) to 1.94 (twice the average length)

Results

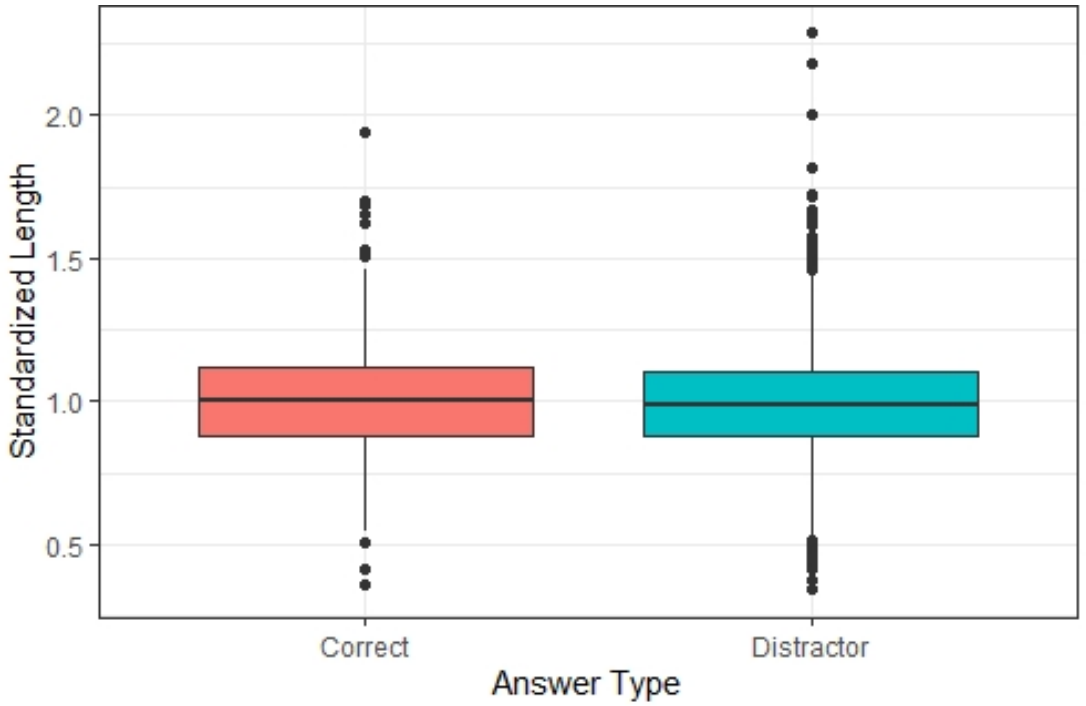


Figure 1

Relative answer choice length by answer type

- There was **no significant difference in standardized answer choice length** between correct (M=1.00, SD=.20) and incorrect (M= 1.00, SD=.21) answer choices ($F(1,1870)$= 0.007, p=.935).
- Variance in answer choice length between the two types of answer choices were not significantly different ($F(1,1870)$= 0.014, p= .905).
- Note: SMEs and item writing staff were trained to keep answer length choices equal

Results

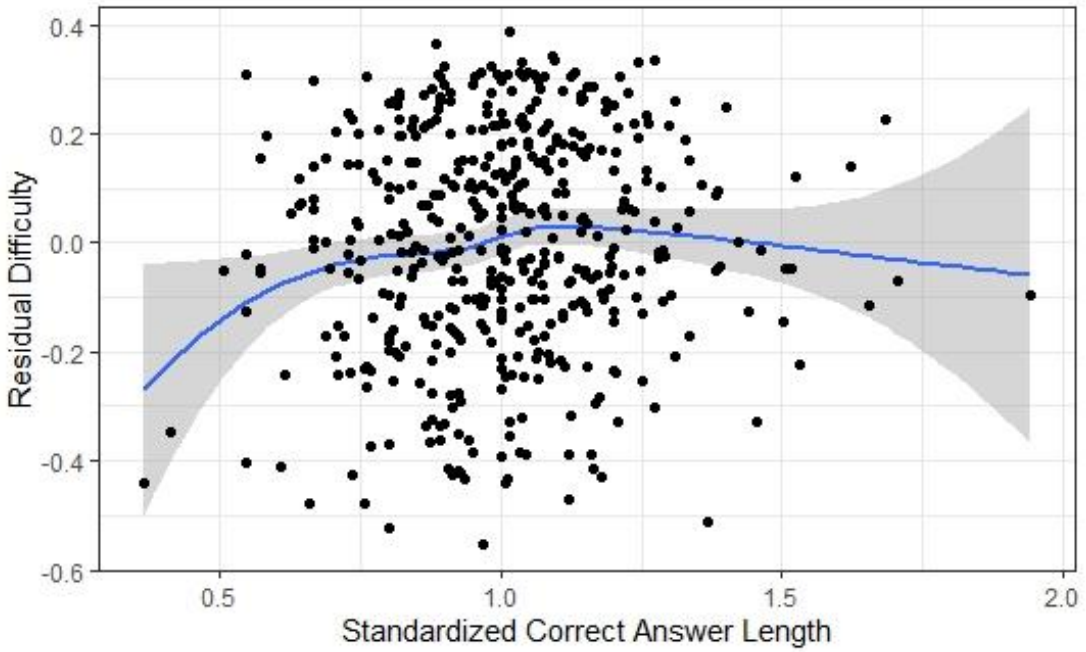


Figure 2

Residual difficulty by relative correct answer choice length

- Relative correct answer length was **not sig. predictive of raw item difficulty** (p=.102).
- After controlling for expected difficulty, regression results indicated a sig. quadratic relationship between relative correct answer length and item difficulty (p= .032)
- When correct answers were **either shorter or longer** relative to their distractors, items tended to be **more difficult than expected based on Angoff/ standard setting ratings**

Results

- Were SMEs **overestimating the impact of uneven answer choice lengths** on item easiness?
  - There was **no relationship** between **relative correct answer choice length and difficulty rating**.
  - Both linear (p=.858) and quadratic (p=.389) relationships were nonsignificant
  - This suggests that the relationship between uneven answer choice length and residual item difficulty is ***not* driven by SMEs overestimating the effect of uneven answer choice lengths**.

Results

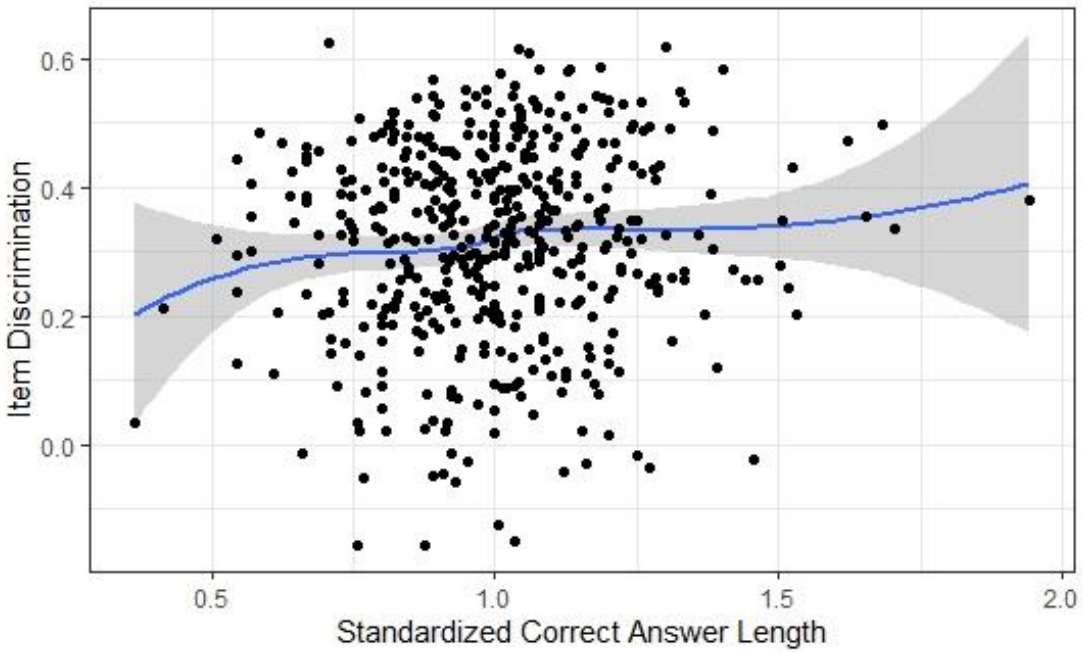


Figure 3

Item discrimination by relative correct answer choice length

- There was **no relationship** between relative **correct answer length and item discrimination**
- Contrary to past research and guidance
  - questions with longer right answers more likely to be guessed correctly

Summary

- RQ1: *How often do correct answer choices differ significantly in length from their distractors?*
  - Efforts to maintain even answer choice lengths were successful; on average, **correct answer choices and distractors were very similar in length.**
  - However, there was some variation in relative correct answer choice length.
- RQ2: *How does the relative length of the correct answer choice correlate with item properties, including a) difficulty and b) point-biserial discrimination?*
  - Items with uneven answer choice lengths were **similar in difficulty to items with even answer choice lengths** and **more difficult than SMEs expected**
  - The relationship between **relative correct answer length** and **item discrimination was not significant.**

Conclusions

- Items with uneven answer choice lengths were **more difficult than SMEs expected.**
  - SME standard setting ratings may have been influenced by relative answer choice lengths
  - Facilitators work with SMEs to balance answer choice length during item writing, so SMEs might have this guideline at the forefront of their mind when assessing item difficulty
  - However, results do not support this hypothesis, suggesting that the effect was driven by differences in actual item difficulty.
- Uneven answer choice lengths were **NOT associated with lowered item discrimination.**

Limitations

- Item writing facilitators' efforts to keep answer choice lengths similar may have minimized variation necessary for analysis; items with glaring discrepancies would have likely been edited in review process
- Modified Angoff standard setting ratings are an imperfect proxy for expected item difficulty
- Results may not generalize outside this setting (end-of-course tests for high school level students)

References

- Board, C., & Whitney, D. R. (1972). The Effect of Selected Poor Item-Writing Practices on Test Difficulty, Reliability and Validity. *Journal of Educational Measurement*, *9*(3), 225–233. https://doi.org/10.1111/j.1745-3984.1972.tb00956.x
- Downing, S. M. (2002). Threats to the Validity of Locally Developed Multiple-Choice Tests in Medical Education: Construct-Irrelevant Variance and Construct Underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235–241.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.
- Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4
- Pham, H., Besanko, J., & Devitt, P. (2018). Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple choice question. *MedEdPublish*, *7*, 225. https://doi.org/10.15694/mep.2018.0000225.1
- Plake, B. S., Cizek, G. J., & Cizek, G. J. (2012). The modified Angoff, extended Angoff, and Yes/No standard setting methods. *Setting Performance Standards. Foundations, Methods, and Innovations*, 181–253.
- Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*(1), 250. https://doi.org/10.1186/s12909-016-0773-3
- Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments: Item-writing flaws and student achievement. *Medical Education*, *42*(2), 198–206. https://doi.org/10.1111/j.1365-2923.2007.02957.x

Questions?