2024 ■ Vol. 20 ■ no. 3

Teaching Quantitative Methods Vignettes

Measuring Pickle Fanaticism: An open dataset for teaching instrument development and reliability

Lindsay J. Alley [a] and Jessica Kay Flake [a]

[a]Department of Psychology, McGill University

Abstract ■ We developed the Pickle Fanaticism Scale and collected pedagogical datasets with the items for the purpose of teaching measurement theory concepts. This paper introduces the datasets and accompanying activities focused on item analysis and reliability. The scale includes items that violate item writing guidelines and others that don't, which enables students to use their problem solving and critical thinking skills to understand how violating these guidelines can impact the psychometric properties of a scale. In our demonstration, we ask students to conduct an item and reliability analysis using R and interpret the results while considering the content of the items. The dataset is free to use, and those teaching measurement topics may use our activities or design their own.

Keywords ■ Measurement theory, instrument development, reliability, item analysis, psychometrics. **Tools** ■ R.

✉ lindsay.alley@mail.mcgill.ca

10.20982/tqmp.20.3.v019

Acting Editor ■ Sébastien Béland (Université de Montréal)

Reviewers ■ One anonymous reviewer

Concept to be presented

In this paper, we introduce a dataset of responses to the Pickle Fanaticism Scale. We created this scale and collected responses to it for the purpose of teaching measurement theory concepts to a class of graduate students and advanced undergraduates, and developed assignments and activities that touch on all five sources of validity evidence described in the *Standards for Educational and Psychological Testing* (AERA et al., 2014): response processes, test content, internal structure, relations to other variables, and consequences of testing. To introduce the dataset, we present example questions from activities on item analysis and reliability. First, we describe the relevant concepts; then we present the Pickle Fanaticism Scale and the item response data; finally, we present examples of assignment questions for teaching item analysis and reliability using the data.

When developing a scale, the goal is to write items that produce reliable and valid scores for their intended use and interpretation (AERA et al., 2014). Item analysis, which includes examining item response distributions, inter-item correlations, and item-total correlations, is a helpful window into how people respond to items. Item response distributions can reveal potential threats to validity. For example, if nearly everyone agrees with an item, it may be a statement of fact and not differentiate peoples' attitudes or opinions. Similarly, items meant to measure the same construct should correlate, so zero or weak correlations can signal that responses are invalid. Additionally, inter-item correlations form the basis of other concepts like reliability and factor analysis. Having students examine patterns of item correlations lays the foundation for understanding these techniques.

The concept of reliability originates from classical test theory (Novick, 1966) and describes the consistency of scores. While there are other metrics of reliability, such as test-retest reliability and interrater reliability, we will focus on internal consistency reliability of a set of items on a scale. According to classical test theory, there are two

sources of variation in item responses: true score variance, and error variance. Each respondent's observed score on the measure, X, can be decomposed into:

$$X = T + E$$

where T is their so-called "true score" on the measure, and E represents the error. According to the classical test theory definition, the error portion of the respondent's score consists of random variations in their responses that would not occur in the same way if they were tested again. For example, when answering a question about how much TV you watch in a week, you might simply forget about a movie night in one moment but remember it if asked the question at a different time. This would change your answer to the question in a random way for reasons unrelated to how much TV you watched. The true score is a hypothetical representation of a person's score on the measure if it were without error. This score decomposition is central to understanding reliability because reliability is a ratio of true score variance to total variance. This provides a representation of how much random error influences observed scores.

One of the most commonly used coefficients of internal consistency reliability is *coefficient alpha* or *Cronbach's alpha* (α; Cronbach, 1951), which can also be estimated as a type of intra-class correlation (ICC). The ICC framework can be expanded to consider reliability across time-points and raters (Yen & Lo, 2002), but here we focus on α, which is a lower bound to the true reliability of a test, so the proportion of true score variance is at least equivalent to α. This coefficient represents the average correlation between all possible halves of the test, and is calculated as follows:

$$\frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sum_{i \neq j} cov_{ij} + \sum \sigma_i^2}\right)$$

where k is the number of items, σ_i^2 is the item variances, and cov_{ij} is the item covariances. This formula shows that α increases when the correlations between items are large, as this will reduce the value of the final fraction. The value of α also increases with the number of items, because the number of elements in $\sum_{i \neq j} cov_{ij}$ increases much more quickly than $\sum \sigma_i^2$ with the addition of more items, which similarly reduces the value of the final fraction.

α is the most used metric of reliability (Flake et al., 2017) and implemented in nearly every software package, which means it is important for students to understand how to use it and evaluate its use. Additionally, its straightforward formula offers insight into the impact of number of items and inter-item correlations on reliability, which scaffolds the understanding of more complex topics in psychometrics. While there is some controversy surrounding the over reliance on α (McNeish, 2018; Raykov & Marcoulides, 2019; Schmidt et al., 2003), it is necessary that students have a solid understanding of this common coefficient.

The Scale and the Dataset

The Pickle Fanaticism Scale includes 33 items that assess three factors: desire to eat, extreme liking of, and desire to evangelize about pickled products.[1] For a complete listing of all items, see Table 1. To give students a clear understanding of the purpose of the scale, which is essential to the task of validation, we constructed the following backstory: market researchers at EvilCorp have been hired by a prominent pickle company to increase sales. Their chosen strategy is to recruit pickle influencers on social media, and they need a measurement tool to identify candidates who are high in pickle fanaticism. They have defined pickle fanaticism as a general zeal for pickled products, and are especially interested in the third candidate factor, the desire to tell others about pickles.

We specifically included items that violated the item writing guidelines in the textbook we used in our measurement theory course (Bandalos, 2018) so that students could practice identifying violations of these guidelines and see for themselves the impact of item wording on psychometric properties. For example, the guidelines state to avoid double-barreled items, use simple vocabulary that respondents will understand, and not to include items that are statements of fact. Can you spot the items that violate these three guidelines?[2]

The data were collected online in September 2021, and 563 respondents were recruited using Twitter (now X). The dataset includes item responses (item1 – item33), calculated total scores for each hypothesized factor, and demographic characteristics, including age, gender, continent where the participant is located, level of English fluency, and education level. The dataset is available in csv and Rds format, along with a codebook, and the course assignments (osf.io/7d9jp).

The Activity: Using the Data for Item and Reliability Analysis

For the assignments, students were asked to complete analyses to help develop a reliable measure of pickle fanaticism with strong validity evidence for its intended use. Students completed five assignments throughout the course,

[1] We acknowledge that this term has religious connotations. Instructors who would prefer to avoid this may change the name of the factor, as none of the items use this term directly.

[2] We have not provided the answer key with this article to maintain the integrity of the assignments but are happy to share it with instructors upon request.

Table 1 ■ Items and Response Options for the *Pickle Fanaticism Scale*

Items				
Strong desire to eat pickled products	**Extreme liking of pickled products**	**Feels the need to evangelize about the benefits of pickles**		
I think about eating pickles at most meals.	Pickles are made with vinegar.	I have many friends who like pickles.		
Over the past 30 years I have eaten thousands of pickles.	I only eat pickles.	Weekly I make sure to tell a friend about pickles, the different kinds of pickles, where you can buy them, and how much they cost.		
I often contemplate the role of pickles in a postmodern society.	I always eat pickles for breakfast.	My family should eat pickles regularly.		
I dream about pickles.	I do not like pickles much.	I don't like my friends who don't eat pickles.		
I prefer to eat other snacks over pickles.	I like pickles less than other foods.	Everyone should eat pickles.		
My go to snack is a pickle.	Pickles taste extremely good.	Pickles are great gifts.		
I would rather eat dill pickles than sweet pickles.	I don't mind pickles.	I like to tell people about my favorite pickles.		
I eat pickles often.	Pickles are a unique food.	I want to share my love of pickles with the world.		
Pickles are delectable sustenance.	Pickles taste bad.	My friends should not eat pickles.		
I would like to eat pickles every day.	Pickles are delicious.	I'm secretive about my pickle habits.		
I avoid eating pickles.	I like pickles a lot.	I recommend pickles to people I know.		
Response Scale				
Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree

and then collated their findings into a full validity report as an end of term project. The item analysis and reliability assignments were the second and fourth assignments in the course, and we describe example questions from each below. All statistical analyses for the course were completed using R.

Item Analysis: Reverse Scoring and Correlation

We instructed students to examine a correlation matrix of the items from the pickle evangelism subscale and then asked them to identify, based on the wordings, whether any items need to be reverse coded. Reverse coding describes reassigning the numerical values for a negatively worded item so that it can be added to the other items to create a meaningful total score (i.e. 1 = 5, 2 = 4, etc.). We focused on one subscale for this assignment to narrow the scope and encourage students to be more detailed and reflective. After students used R to reverse code items 31 ("My friends should not eat pickles") and 32 ("I'm secretive about my pickle habits"), they produced the correlation matrix (see Figure 1).

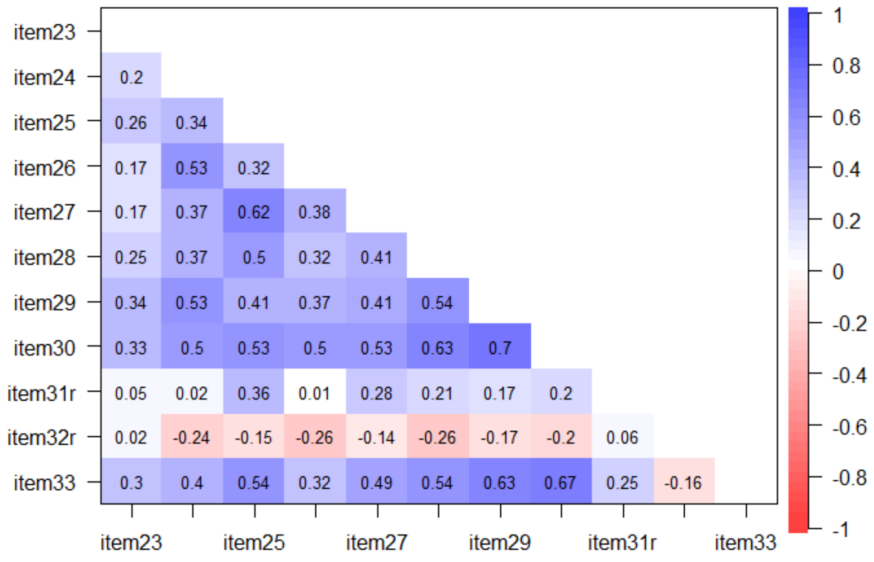
After reverse coding item 32, it should be positively correlated to the other items, but instead becomes negatively correlated. Students were asked to interpret this unexpected result and consider how the item content contributes to the negative correlations. Reverse scoring assumes that the negatively worded item is just the opposite of the positive, and this example inspires students to consider that pickle secretiveness may not be the opposite of pickle evangelism. It may be the case that people who like to recommend pickles to others, and give pickles as gifts, are also secretive about how often they eat pickles.

Reliability: Correlation and Number of Items

The reliability assignment was the fourth out of five assignments. The assignments build on each other towards the goal of developing the final Pickle Fanaticism Scale, and many items had been removed before this assignment based on evidence from think-aloud interviews, item analysis, and factor analysis. The final selected items loaded onto two factors: liking pickles (6 items), and pickle evangelizing (5 items).

Students first calculated α and its confidence interval for each factor using the alpha function from the `psych`

Figure 1 ■ Correlation Matrix of Items. This correlation matrix was produced after reverse coding `item31r` and `item32r`.



package in R (Revelle, 2024). This function has many other useful outputs related to psychometrics, such as the item-total correlations, and the reliability of the scale if that item was removed. To ensure that students understood how to interpret this output and use it to make decisions, we asked them which item they would remove from each factor and to specify the part of the output they used to make their decision. While achieving adequate reliability is important, over-emphasis on this metric has negative consequences for the validity of measures (Clifton, 2020). To guard against this, we asked students whether they would still remove the two items if they considered all evidence gathered in their assignments so far, and to explain why or why not.

Conclusion

Psychometrics and measurement theory can be challenging to teach because these topics are highly technical and quantitative, but also need to be successfully connected to validity theory and the qualitative aspects of item writing. A dataset of real responses to items is a rich resource for making those connections. Students can experience how item meaning and wording impact participants' responses, and how that is visible in the psychometric results, including item statistics and reliability. In addition to the activities described in this paper, we used the Pickle Fanaticism Scale to teach validity evidence sources as described in the *Standards* (AERA et al., 2014): response processes, test content, internal structure (i.e., Exploratory Factor Analysis), relations to other variables (using a second dataset with other scales included), and consequences of testing. Instructions for these assignments are included in our online repository. As well as using our materials, we encourage others to develop and share their own activities using the data.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association. https://play.google.com/store/books/details?id=clI_mAEACAAJ

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259–270. doi: 10.1037/met0000236.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi: 10.1007/bf02310555.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. doi: 10.1177/1948550617693063.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. doi: 10.1037/met0000144.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. doi: 10.1016/0022-2496(66)90002-2.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, *79*(1), 200–210. doi: 10.1177/0013164417725127.

Revelle, W. (2024). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.4.1]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*(2), 206–224. doi: 10.1037/1082-989x.8.2.206.

Yen, M., & Lo, L.-H. (2002). Examining test-retest reliability: An intra-class correlation approach. *Nursing Research*, *51*(1), 59–62. doi: 10.1097/00006199-200201000-00009.

Citation

Alley, L. J., & Flake, J. K. (2024). Measuring Pickle Fanaticism: An open dataset for teaching instrument development and reliability. *The Quantitative Methods for Psychology*, *20*(3), v19–v23. doi: 10.20982/tqmp.20.3.v019.

Received: 15/02/2024 ~ Accepted: 09/05/2024

Extended activity metadata

Concept illustrated	Item analysis and reliability	*Type of activity*	Take home assignment
Prerequisite	Correlation, confidence intervals, statistical graphics	*Types of data*	Survey data
Co-requisite	Statistical programming, validity theory	*Computation by*	R
Suitable class size	20-30 students	*Duration*	2-3 hours per assignment