34th Australasian Association for Engineering Education Conference
3 - 6 December 2023 – Gold Coast, QLD

# Improving data collection: A comparative analysis of criterion-based interpretation framework items with Likert items based on procedural justice and team formation in engineering

Jolanta Szymakowski[a]; Nicoleta Maynard[a], and Callum Kimpton[a].
*Monash University[a]*
*Corresponding Author Email: yola.szymakowski@monash.edu*

## ABSTRACT

### CONTEXT

Developing questionnaires to collect data is a common method is educational research. People's perceptions are typically garnered using Likert-based questionnaire items, where respondents indicate their level of agreement to statements. However, items that are not-Likert based can also be used. An alternative approach is for item responses not to describe levels of agreement but instead to describe behaviours in what is known as a criterion-referenced interpretation framework. This paper compares both ways of generating items and item responses for the same latent construct, procedural justice. Procedural justice is considered one important lens in team formation, especially when engineering students form teams as part of their studies.

### PURPOSE OR GOAL

The purpose of this study is to conduct a comparative analysis of two questionnaires used to evaluate a specific facet of team formation, procedural justice, with the aim to demonstrate the potential for improved questionnaire performance through the implementation of recent advances in quantitative measurement research. By providing examples of alternative approaches to item presentation, this research seeks to facilitate the development of more effective questionnaires by educators/researchers in the engineering field.

### APPROACH OR METHODOLOGY/METHODS

The paper is of a theoretical or methodological nature and therefore does not involve the collection of empirical data. It presents a comparative analysis and discussion of two questionnaires, one utilising Likert-type items and the other utilising criterion-referenced items.

### ACTUAL OR ANTICIPATED OUTCOMES

An introduction to a criterion-based approach to writing a questionnaire will be described, so that engineering educators (i) understand some of the latest developments in measurement; (ii) have exposure to concepts such as Rasch and Item Response Theory, and (iii) have the confidence to write non-Likert items.

### CONCLUSIONS/RECOMMENDATIONS/SUMMARY

Questionnaires are widely used to gather data, and Likert based items are popular. However, questionnaires that are criterion-referenced and use Item Response Theory produce sound measurement instruments and are not overly challenging to construct. By conducting a comparative analysis of two questionnaires assessing the same latent construct this research sheds practical light on these concepts for engineering educators.

### KEYWORDS

Criterion-referenced assessment, Item Response Theory, procedural justice.

# Introduction

Questionnaires are a method of data collection that is extensively used in educational research. The data collected through a questionnaire typically meet one of three research needs: (i) to gauge the attitudes or perceptions of the respondent about the idea presented in the item stem, often looking for statistical patterns with other variables; (ii) to survey a large number of people and determine the state of play of the population landscape; and (iii) to create a measurement instrument. Questionnaire items thus need to be tailored to meet the different research needs.

The first research need is based on determining people's attitudes and perceptions, and item responses often adopt a Likert approach. People may be asked to rate their level of agreement with the idea presented in the item stem, with the responses typically coded 5 (strongly agree), 4 (agree), 3 (neither agree nor disagree), 2 (disagree) or 1 (strongly disagree). Means of groups of items are determined, and statistical patterns with other variables often calculated. For example, an analysis of the correlations between personality traits, teamwork competencies and academic performance among first year Malaysian engineering students used Likert items to quantify personality traits and Likert items to quantify teamwork competencies, as presented in ITPmetrics.com (Tang, 2021). Statistical analyses such as correlations and regressions linking Likert responses with academic performance were undertaken with a view to determining significant patterns with academic achievement.

The second research need is to survey a large number of people to determine the state of play, with a census such as the Australian census being an example of this. Questionnaire items address not attitudes nor perceptions, but concrete numerical data, for example, asking about a person's income with item responses requiring the respondent to choose one of a number of possible income ranges. The data sought are counts and frequencies, so that trends are noted. The census data help inform policy in economic, social, population and environmental matters of importance to policymakers (Australian Bureau of Statistics, 2023). Subsets of census data may be used to note trends in engineering education and engineering practice (e.g., Palmer & Campbell, 2016).

The third research need that the collection of questionnaire data meets is to gather data so that a measurement instrument may be developed. Measurement has a precise definition (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014) and there is a recommenced approach to developing a valid measurement instrument (American Educational Research Association et al., 2014; Wolfe & Smith, 2007). Questionnaire items may be Likert, multiple choice, or criterion referenced. In a criterion referenced interpretation framework, item responses are descriptions of behaviours ranging from a low level of the behaviour to increasingly higher levels of the behaviour (Griffin, 2014; Griffin & Gillis, 2000). After item responses are coded, they are scaled according to a scaling model. Rasch models, for example, take the item responses which are ordinal, transforming them into integer responses, "mapped onto a continuum that represents measurable quantities of the target construct" (Wolfe & Smith, 2007, p. 220). The outcome of a scaling model is a measure based on integer data.

Questionnaires are used to collect data that are used to quantify a concept. However, merely assigning a number to an object is not the same as measuring it (Wu, Tam, & Jen, 2016). Psychometric theory has evolved a more nuanced definition of measurement, based on the idea of a ruler or thermometer: "Measurement begins with the idea of a variable or line along which objects can be positioned, and the intention to mark off this line in equal units so that distances between points on the line can be compared" (Wright & Masters, 1982, p. 1). Measures allow for comparisons between respondents. This necessitates an additional step in data analysis – scaling.

 

## Aims and Objectives

Educational researchers often use an existing questionnaire when studying student behaviour. However, a claim that the questionnaire 'measures' a construct may be premature. This paper takes a concept used in the formation of engineering student teams, procedural justice, and compares the measurement approaches using questionnaire items based on the first and third research needs showing differences between what the quantified data can reveal. Items are taken from a questionnaire developed to measure procedural justice. Procedural justice is a concept from organisational psychology that focuses on decisions that are made that impact people in an organisation. Procedural justice is the perceived fairness of the process by which those outcomes were arrived at (Cohen-Charash & Spector, 2001). The sense of procedural justice is predicted to be related to cognitive, affective, and behavioural reactions toward the organisation, such as organisational commitment (Murphy & Tyler, 2008).

In this research procedural justice is applied not in an organisational content but to students as they form teams in group-work based projects. The decisions that are made being about them are both their allocation to groups, as well as, once the groups are formed, decisions about grades and procedures set up within the groups to handle conflict and workload allocation. The students were in the first year of an engineering curriculum which was based on a common first year, common to all offered engineering disciplines at one Australian university.

## Procedural Justice

The construct procedural justice is theorised to consist of 8 aspects: Process Control, Decision Control, Consistency, Bias Suppression, Accuracy, Correctability, Representativeness, and Ethicality (Colquitt & Rodell, 2015). In this paper, the first aspect, Process Control, is considered. Process control is about voice, where procedures provide opportunities for the people affected by those procedures to have a voice. A procedure that allows participants to have a voice will be considered fairer than a procedure that does not allow participants to have their say (Colquitt, Greenberg, & Zapata-Phelan, 2005; Donia, Mach, O'Neill, & Brutus, 2022).

Table 1 presents the items used to collect data about Process Control from two different questionnaires. The first questionnaire, the Likert based Colquitt and Rodell (2015) measure, presents one item that taps into process control, asking about participants' perceptions about having a voice. The second questionnaire, developed for this research, presents two items about process control. The teams in the items are student teams, and managers is the term used to describe the tutors to which a number of student teams have been allocated. In contrast to the Likert item, the criterion referenced responses are descriptions of behaviours, or what is done, said, made, or written (Griffin, 2014; Robertson et al., 2022).

 

**Table 1 Process Control Questionnaire Items**

| | Process Control – Likert (Colquitt & Rodell, 2015) | Process Control - criterion referenced interpretation framework |
|---|---|---|
| 1 | The question below refers to the procedures your supervisor uses to makes decisions about group communication.<br>To what extent are you able to express your views during those procedures?<br>1 = To a Very Small Extent,<br>2 = To a Small Extent,<br>3 = To a Moderate Extent,<br>4 = To a Large Extent,<br>5 = To a Very Large Extent. | During team meetings, team members ...<br>1.2 All speak to discuss and raise issues affecting the team.<br>1.1 Tend to leave the discussions to one or two team members, although the quiet team members tend to speak with others outside the team.<br>1.0 Tend not to speak or raise issues, or only one person tends to speak. |
| 2 | | Managers and unit coordinators are available to be respond to questions and concerns in the online forum, face-to-face or via email. Managers and unit coordinators typically respond ...<br>2.2 Within 24 hours.<br>2.1 Within 72 hours.<br>2.0 In timeframes that are not useful. |

# Discussion

Likert responses and criterion referenced responses undergo different data analyses, as described in the following sections.

## Data Analysis – Likert

Typically, Likert responses are coded as integer data – 'to a very small extent' coded as 1, 'to a small extent' coded as 2, etc, as shown in Table 1. The interval between each response is taken to be the same (i.e., 1), as well as assuming that the value of a 'to a very small extent' for every item in the questionnaire is the same (i.e., 1). There are documented procedures to determine averages for the data and correlates with other variables (e.g., Pallant, 2011). An average of 3.2 might be reported as the outcome of item 1, indicating respondents' agreement is more on the large extent side than the small extent side. In Tang's example (2021), respondents answered 120 Likert questions to cover all five personality dimensions, with each personality trait described by one number—the mean of the typically 24 responses for each trait. This one number was used in further analyses with other variables. This one number is in essence 'point data'.

More nuanced thinking challenges the assumption that it is appropriate to code the ordinal data of a Likert response as integer data (Bond, Yan, & Heene, 2020; Stevens, 1946). Appropriate data analyses for ordinal data include summaries of frequencies, mode, median, and ranges. Means and standard deviations apply just to integer and ratio data, yet it is common to see Likert data invalidly described with means and standard deviations.

However, Likert responses can be converted into integer data by an additional step, scaling (Bond et al., 2020; Wolfe & Smith, 2007). Scaling is a process where the ordinal codes "are

 

combined and mapped onto a continuum that represents measurable quantities of the target construct" (Wolfe & Smith, 2007, p. 220). An example of scaling is described in the next section.

## Data Analysis - Criterion Referenced Interpretation Framework

Under a criterion-referenced interpretation framework, responses are descriptions of what is seen, done, made, or written, reflecting behaviours from a lower to higher levels of the construct. The item response describing the lowest level of behaviour is coded as 0, the next highest is coded 1 and so on to the highest level of behaviour described in the item. These numbers are just labels and only serve to order the item responses in preparation for scaling. The Rasch scaling model used in this instrument used to measure procedural justice is the Partial Credit Model, as one or more levels of performance can be identified and each item may have its own unique scoring code (Rasch, 1980; Wolfe & Smith, 2007).

Scaling software include WinSteps (Linacre, 2023), the R package TAM (Robitzsch, Kiefer, & Wu, 2022), and ConQuest (Adams, Wu, & Wilson, 2015). Figure 1 shows one of the outputs of scaling, the Wright map. The Wright map shows the ruler or thermometer of the construct with person ability (the xs on the left) and item difficulty (item responses on the right) on the same scale. All 16 items in the questionnaire are shown, but the responses for items 1 and 2 are highlighted. The scaling software does not present the bottom response (1.0 and 2.0).

The unit of measurement for scales in Item Response Theory is the logit, which is short hand for taking the 'logarithm of it' (Andrich & Marais, 2019), or the 'log of odds unit' (Wu et al., 2016). Ability and item difficulty are shown on the same scale on the Wright map, and the distance between a person's ability and the item difficulty on the scale is defined as the logarithm of the odds of success of the person on the item, where odds is the ratio of the probability of success over the probability of failure. A logit is thus defined as $\ln(p/(1-p))$, where p is the probability of success (Wu et al., 2016).

Figure 1 shows that response 1.1 measures - 1.19 logits on the procedural justice scale, response 1.2 is at 0.37 logits, a difference not of 1 as assumed in the Likert scale, but of 1.56; 2.1 is at -0.48 logits, and 2.2 is at 0.51 logits, a difference of 0.99 logits. For this particular group of students, when it comes to measuring procedural justice, students experienced low to middling levels of procedural justice. The developed ruler or measure reflects this group of respondent's experiences.

Scaling generates a ruler which allows comparisons among different groups of respondents. It says that, for this group of respondents, some respondents experienced a high amount of procedural justice, and that the questionnaire could be improved by adding additional items above 11.2 to tap into higher levels of procedural justice. This provides a way forward to develop a more nuanced understanding of the construct being measured, procedural justice.

 

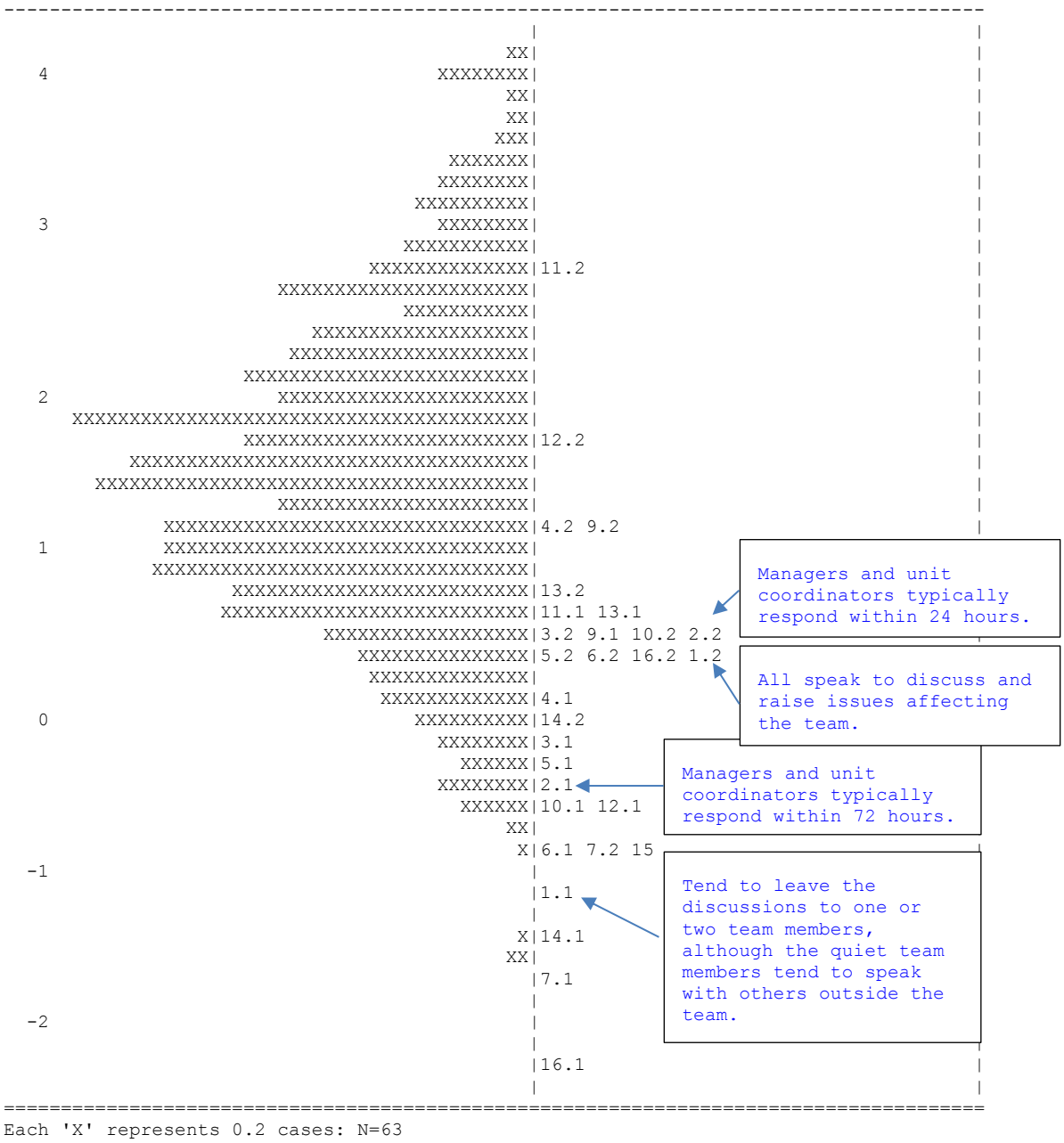


**Figure 1 Wright map for procedural justice**

## Item Development Notes

When developing items in a criterion-based reference framework, there is a risk that item responses may describe a behaviour that the respondent has not experienced or miss describing a behaviour. The process for developing items and item responses relies on expert input and pilot studies (Griffin & Gillis, 2000; Wolfe & Smith, 2007) to address this, but the items may need to be modified in the light of feedback from respondents. Additional item responses may be added, including the equivalent of a n/a where appropriate.

## Developing a questionnaire with criterion referenced items

Wolfe and Smith (2007) describe the clearest process for describing how to construct a questionnaire that gathers data that creates a valid measurement instrument. The creation of

criterion referenced items is described in Griffin's Assessment for Teaching (2014) and Robertson's et al Writing Objective and Judgement-based Assessment Items (2022).

# Conclusion

When understanding human behaviour, questionnaires are used to gather data. The data is then analysed into scales, often stating that the questionnaire 'measures' the construct. However, measurement is more than just assigning numbers to objects (Wu et al., 2016). More nuanced definitions of measurement state that "Measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes" (Nunally & Bernstein, 1994, p. 1) and "Measurement begins with the idea of a variable or line along which objects can be positioned, and the intention to mark off this line in equal units so that distances between points on the line can be compared" (Wright & Masters, 1982, p. 1). These definitions move beyond simply the allocation of numbers to the creation of a ruler or a thermometer. The ruler can be created by appropriate scaling of the responses.

If your research questions are based around people's attitudes and perceptions at a point in time, then Likert responses are appropriate. If your research questions seek to compare groups of people on a behaviour, then developing a measurement instrument is apt. When measuring behaviour, describing those behaviours using criterion-referenced items is fitting.

For engineering education researchers who use an older instrument or who design their own instrument to 'measure' an aspect of student behaviour, it is important to recognise that psychometricians have developed standards and more nuanced understandings of measurement. Questionnaires are an effective data gathering tool, but clarifying the purpose of the questionnaire helps determine the types of items to be written, and appropriate data analysis. Likert responses described by summary statistics may not be a sophisticated enough approach to answer your research questions, and other approaches are now more readily available.

## References


Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software (Version 4) [Software]. Camberwell, Victoria: Australian Council for Educational Research. Retrieved from https://www.acer.org/conquest

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: AERA.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer. https://doi.org/10.1007/978-981-13-7496-8

Australian Bureau of Statistics. (2023). Value of the Census | Australian Bureau of Statistics. Retrieved from https://www.abs.gov.au/census/about-census/value-census

Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the Rasch Model: Fundamental measurement in the human sciences (4th ed.). New York: Routledge. Retrieved from https://doi.org/10.4324/9780429030499

Cohen-Charash, Y., & Spector, P. E. (2001). The role of justice in organizations: A meta-analysis. *Organizational Behavior and Human Decision Processes, 86*(2), 278-321. 10.1006/obhd.2001.2958

Colquitt, J. A., Greenberg, J., & Zapata-Phelan, C. P. (2005). What is organizational justice? A historical overview. In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational*


 

*justice* (pp. 3-56). London, UNITED KINGDOM: Psychology Press. https://doi.org/10.4324/9780203774847

Colquitt, J. A., & Rodell, J. B. (2015). Measuring justice and fairness. In R. S. Cropanzano & M. L. Ambrose (Eds.), *Oxford Library of Psychology*. *The Oxford handbook of justice in the workplace* (pp. 187-202). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199981410.013.0008

Donia, M. B. L., Mach, M., O'Neill, T. A., & Brutus, S. (2022). Student satisfaction with use of an online peer feedback system. *Assessment & Evaluation in Higher Education, 47*(2), 269-283. 10.1080/02602938.2021.1912286

Griffin, P. (Ed.). (2014). *Assessment for teaching*. Port Melbourne, Victoria: Cambridge University Press, 2014.

Griffin, P., & Gillis, S. (2000). Valuing Vocational Education & Training Outputs *Standardised Output Measure: Pairwise Comparison Project*, Assessment Research Centre: University of Melbourne.

Linacre, J. M. (2023). Winsteps [Software]. Retrieved from https://www.winsteps.com/index.htm

Murphy, K., & Tyler, T. (2008). Procedural justice and compliance behaviour: the mediating role of emotions. *European Journal of Social Psychology, 38*(4), 652-668. 10.1002/ejsp.502

Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Pallant, J. (2011). *SPSS survival manual: A step by step guide to data analysis using SPSS* (4th ed.). Crows Nest, NSW: Allen & Unwin. Retrieved from https://doi.org/10.4324/9781003117452

Palmer, S., & Campbell, M. (2016). *Characterising the Australian engineering workforce and engineering graduate occupational outcomes using national census data*. Paper presented at the PAEE/ALE 2016 : Proceedings of the 8th International Symposium on Project Approaches in Engineering Education and 14th Active Learning in Engineering Education Combined Conference and Workshop, Guimarães, Portugal. Retrieved from http://hdl.handle.net/10536/DRO/DU:30084975

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Chicago: The University of Chicago Press.

Robertson, P., Beswick, B., English, N., Kheang, T., Collins, M., Nguyen, C., . . . Awwal, N. (2022). Writing objective and judgement-based assessment items: Assessment Research Centre, The University of Melbourne. https://doi.org/10.26188/20479536

Robitzsch, A., Kiefer, T., & Wu, M. (2022). TAM: Test Analysis Modules [R package version 4.1-4]. Retrieved from https://cran.r-project.org/web/packages/TAM/index.html

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680. Retrieved from https://www.jstor.org/stable/1671815

Tang, K. H. D. (2021). Personality traits, teamwork competencies and academic performance among first-year engineering students. *Higher Education, Skills and Work-Based Learning, 11*(2), 367-385. 10.1108/HESWBL-11-2019-0153

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 202-242). Maple Grove, Minnesota: JAM Press.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer. 10.1007/978-981-10-3302-5

## Copyright statement