Original Research

SAGE Open
October-December 2024: 1–18

DOI: 10.1177/21582440241295501
journals.sagepub.com/home/sgo
Sage

Not Liking the Likert? A Rasch Analysis of Forced-choice Format and Usefulness in Survey Design

Celeste Combrinck[1]

Abstract

We have less time and focus than ever before, while the demand for attention is increasing. Therefore, it is no surprise that when answering questionnaires, we often choose to strongly agree or be neutral, producing problematic and unusable data. The current study investigated forced-choice (ipsative) format compared to the same questions on a Likert-type as a viable alternative. An established motivation questionnaire was administered in two versions, forced-choice and Likert-type, to 1088 first-year engineering students. Descriptive, non-parametric statistics and Rasch measurement models were applied to assess usefulness, validity and reliability. Results: The ipsative version had a higher response rate, less missing data, and the motivations emerged more clearly. Evidence for the reliability and validity of the forced-choice version was excellent. The forced-choice format is recommended as an alternative to the Likert types when collecting human or social survey data.

Plain language summary

Ask participants to choose their top options in surveys—How to use and defend forced-choice items

Questionnaires can be great tools for collecting data but often require time from respondents to read and rate each statement. The current study used a forced-choice format, where respondents had to choose their top reasons, as an example of how to collect data more efficiently. Asking respondents to choose their preferred options or rank questions can also lead to more precise answers. When you give respondents a rating scale, they tend to opt for the middle choice or agree with most statements. Asking them to choose their preferred option instead took less time and concentration, resulting in more people completing the questionnaire and better-quality data. The current article demonstrates how to evaluate the consistency and legitimacy of the questionnaire with the Rasch model. The article argues for forced-choice formats and applying a psychometric theory so that researchers can show evidence for the measurement quality of their instrument and obtain valuable results.

Keywords

forced-choice format, Rasch measurement models, Ipsative data, Likert-type format, reliability and validity of survey data

Introduction

Rensis Likert originally designed his rating scales to reflect the underlying construct through the level of agreement on a numerical continuum (Likert, 1974). His original idea was that the level of agreement could be summed or averaged and that each item contributes to the measured central dimension (H. Boone & Boone, 2012). The assumption that aggregated responses represent the underlying construct is based on Classical Test Theory (C.T.T.) (South et al., 2022; Warmbrod, 2014). Critique of the Likert-type scale includes the assumption

[1]University of Pretoria, Gauteng, South Africa

Corresponding Author:
Celeste Combrinck, Science, Mathematics and Technology Education (SMTE), Faculty of Education, University of Pretoria, Room 4-3 Natural Sciences Building, Pretoria, Gauteng 0001, South Africa.
Email: celeste.combrinck@up.ac.za

Data Availability Statement included at the end of the article

that they lead to interval variables, whether parametric tests are appropriate for such formats and the oversimplification of complex attitudes created from the response structure (Harpe, 2015; Sullivan & Artino, 2013). The other critique of Likert scales is that they can lead to response biases such as social desirability and central tendency responding (Kusmaryono et al., 2022), addressed in the current article. Self-report data of any type can be vulnerable to acquiescence bias, social desirability and neutral responding, which Likert-type response options can potentially aggravate (Brown & Maydeu-Olivares, 2018a, 2018b; Combrinck & Inglis, 2020; Wang et al., 2017). The interpretation of Likert options and ratings may also vary among respondents, causing other potential biases due to the ambiguities of language (Hancock & Volante, 2020). Using a questionnaire which has shown acceptable and high-reliability indices will not necessarily reduce over-reporting of "good" behavior, and the halo effect might persist (Douglas & Tramonte, 2015; Kreitchmann et al., 2019). Respondents might be tempted to endorse statements they perceive as socially acceptable (Latkin et al., 2017; Vesely & Klöckner, 2020). Careless or inattentive survey responses could be due to several factors, including poor targeting, inappropriate sampling or poor item writing (Jaeger & Cardello, 2022). Berry et al. (2019) found that inattentive responses correlate with being male, low on personality factors such as conscientiousness and agreeableness, and less sensitive themes included in the items (whereas susceptible topics, e.g., crime, drew more attention from respondents). The number of categories in a Likert scale could increase inattentive responding, and M. A. Revilla et al. (2014) found that five categories were the maximum number that should be employed. Even transient psychological states (e.g., negative affect) could influence a respondent (Huang & Wang, 2021). While we have little to no control over the respondent's traits and feelings when completing a survey, the instrument's quality is the researcher's responsibility. This is why using Rasch models is strongly advocated to assess any format type of instrument, including Likert-types (Knoch & McNamara, 2015; Retief et al., 2013). From a measurement perspective, many of these critiques leveled at Likert formats can be addressed by testing the hypothesis that respondent ability to endorse more of the construct aligns with the difficulty of item endorsability, that is, applying the Rasch rating scale model, and testing the claim that an underlying construct is consistently being represented by all items (Combrinck, 2020; Fisher, 2009). The researcher should be aware of the shortcomings of using any particular format and examine underlying assumptions of measurement (Fisher, 2022). Likert item types remain immensely popular in survey design due to familiarity with the format and ease of analysis.

Alternative formats have been suggested, such as dichotomous scales (Dolnicar et al., 2011), fuzzy rating scales (Castaño et al., 2020), slider scales (Kemper et al., 2020) and rankings (Yannakakis & Martínez, 2015). Forced-choice questions present the respondents with a list of statements or items, and they choose the most suitable options (Brown & Maydeu-Olivares, 2018a, 2018b). Compared to Likert-style formats, the advantages of forced-choice questions have been explored in various studies with potential advantages well explored (Chan, 2003; Cheung & Chan, 2002; Lee et al., 2019). The Forced-multiple choice format is popular in psychological research, especially personality and career assessments (Brown & Maydeu-Olivares, 2013). Forced-choice formats have disadvantages, such as the complexity of administration and analysis. Administering forced-choice would be especially difficult if a paper and pen route is followed, as the respondents may not understand what is required to answer the question (Buchanan & Morrison, 1985; Zhang et al., 2020). Analyzing traditional Likert-type scales is also more accessible, with established techniques and parametric analysis options after applying logarithmic transformations (Hontangas et al., 2015; Smyth et al., 2006). Deriving interpretations from forced-choice formats can be complicated, making it challenging to compare groups or longitudinal assessments, and this places an additional burden on the researcher (Salgado et al., 2015; van Eijnatten et al., 2015). Some constructs and ratings do not work well with forced-choice formats, and here, the Likert type is preferred as it is more versatile (Lee et al., 2019; J. D. Miller et al., 2018).

Bäckström and Björklund (2024) found that Likert type items and forced-choice options yield similar information. A mixture of items format solution which leaves a backdoor for the researchers could be very beneficial and should be considered (Schulte et al., 2021). Using forced-choice might be valuable in certain instances, for example, constructs that lead to mid-choice favoring (Nadler et al., 2015). At the same time, using Likert-type formats may be advantageous when a response is required for each item, when respondents are more familiar with the format and when analysis requires the numerical range of the items for analysis and comparison purposes (Hall et al., 2016; Nemoto & Beglar, 2014; Wu & Leung, 2017). Despite the advantages of Likert formats, social desirability remains a serious concern. Therefore, the paper presents a case for forced-choice formats in social and human science settings to offer more usable data. Forced-choice questions also have a definitive practical advantage—the format requires less time and can be less attention-intensive for respondents than rating each statement on a Likert-type (Brown,

2016; Brown & Maydeu-Olivares, 2011). The short attention span of humans has been well documented, and we researchers want accurate answers, which our respondents are more likely to give if we do not overburden them (M. Revilla & Ochoa, 2017). The usefulness of item response models (I.R.T.), including the Rasch model, for analyzing ipsative data is well documented (Brown & Maydeu-Olivares, 2013, 2018a, 2018b). Test-retest and comparative studies using Likert-type items have been used to validate ipsative results (Calderón Carvajal et al., 2021). However, the current paper argues that this is unnecessary when using log-transformed measures such as those produced by the Rasch model (Rasch, 1960, 1993; van Alphen et al., 1994). A valid and reliable representation of a construct requires a range of relevant aspects to be examined, and Rasch models offer strategies to evaluate instrument functioning (Bond et al., 2021; W. J. Boone, 2016). Instruments are valid if their results inform decisions, changes and growth (W. J. Boone et al., 2014). When responses lack variance, that is, skewed values, the data quality is severely impacted, and limited inferences can be drawn (Kreitchmann et al., 2019; Xiao et al., 2017). In the current study, I demonstrate how the Rasch model can be used to check the functioning of forced-choice questionnaires, which could lead to increased awareness and application of the format to gain more valuable inferences and enhance meaningful measurement. I offer guidelines for researchers who want to use alternative formats. In previous studies, Rasch models have been applied to ipsative data with promising results (Andrich, 1989; Van Zile-Tamsen, 2017; Wang et al., 2017).

Research Questions

Research question 1: *What does the evidence show us about forced-choice questions as viable alternatives to Likert-types?*

Research question 2: *How can we use Rasch models to evaluate the psychometric properties of forced-choice questions?*

Research question 3: *What does a Rasch analysis reveal about the usefulness of forced-choice questions compared to Likert-types?*

Materials and Methods

Instrument

The *Academic Pathways of People Learning Engineering Survey* (APPLES) contains 15 motivation-related items used in higher education to understand students' drive to study engineering. The questionnaire demonstrated reliability and validity in previous uses in the United Kingdom (Sheppard et al., 2010) and has been used in the United States and South Africa (Direito et al., 2019; Donaldson et al., 2008; Eris et al., 2010). The motivation questionnaire covers six dimensions of motivation:

1. *financial (F)—Example of item: An engineering degree will guarantee me a job when I graduate*
2. *intrinsic behavioral (I.B.)—Example of item: I like to figure out how things work*
3. *intrinsic psychological (I.P.)—Example of item: I feel good when I am doing engineering*
4. *mentor influence (M)—Example of item: A mentor has introduced me to people and opportunities in engineering*
5. *parental influence (P)—Example of item: My parents want me to be an engineer*
6. *social good (S.G.)—Example of item: Technology plays an important role in solving society's problems*

Figures 1 and 2 show the two forms of the APPLES questionnaire administered in this study.

Version 1 is the forced-choice type, where students had to choose up to five reasons for studying engineering. Version 2 is the original Likert-type scale, where a respondent could indicate that the option was *not a reason*, *a minimal reason*, *a moderate* or *a major reason* for studying engineering. Most respondents answered the forced-choice version of the assessment, but a randomly selected subsample ($n = 221$) received the original Likert version for comparison.

Sample

The APPLES questionnaires were administered online via Qualtrics and completed by 1089 first-year South African engineering students who participated in a broader study on student success (Inglis et al., 2022; Inglis & Simpson, 2023). There were more male (71%) than female respondents; almost two-thirds of the respondents were White (58%), followed by African (26%) and other ethnicities such Indian or Mixed race (16%). Most students had a family member who had attended university (81%), and the average age was 19. The sample size is adequate as 68% of first-year students responded to the survey and allowed their data to be used for research purposes. I intend to use non-parametric statistics in future analyses to detect differences in response patterns. Non-parametric tests have fewer assumptions, and my large sample size should enhance the detection of effect sizes and significance. The responses are interdependent, and choosing one option might influence one's choice to select another, but the large sample size should help account for this. Sample size recommendations for Rasch Analysis suggest that complex models require

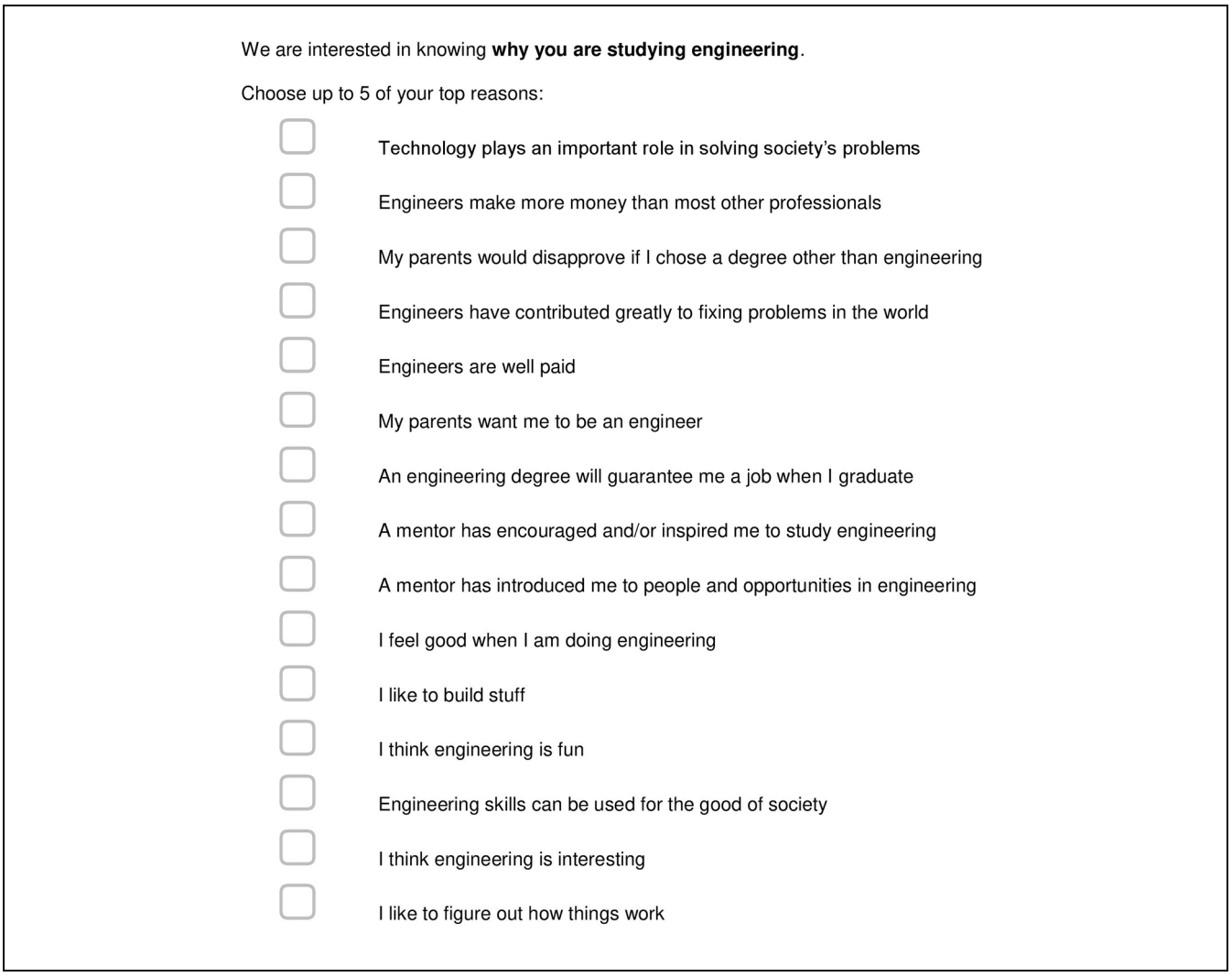

Figure 1. Forced-choice format of the APPLES questionnaire.

larger sample sizes for stability (Linacre, 1994), and the forced-choice nature of the ipsative data increases the complexity. Studies of Rasch parameter stability suggest that sample sizes larger than 250 are adequate for 15 items (Chen et al., 2014).

The Rasch Rating Scale Model (R.S.M.)

The Rasch Rating Scale (R.S.M.) model was designed to analyze survey-type data where the same response options are present, using the Rasch logic of comparing a person's ability to item endorsability (Andrich, 2016). The R.S.M. is uniquely suited to analyzing survey data, especially ipsative data, as each option is evaluated in relation to the other, effectively handling the nature of the comparisons (Engelhard, 2013; Rost, 2001; Vidotto et al., 2018). The R.S.M. comparisons consider the options' interdependencies rather than evaluating items in isolation. The Rasch model also transforms raw scores into logic, a logarithmic scale that converts ordinal data into actual interval scale data. The logit scale can range from negative to positive infinity (although most applications see values within about -3 to $+3$) (Bond et al., 2021; W. J. Boone et al., 2014). The mean item location is usually set to zero logits as part of the estimation process, and the Standard Deviation (S.D.) is sample-specific.

Data Analysis

I applied the Rasch dichotomous model using Winsteps© 5.4.0.0 (Linacre, 2023a). SPSS version 28 was used to generate descriptive statistics and compare dimensions with non-parametric statistics (IBM, 2023). Every question was conceptualized as an option. Students were required to select their most relevant 5 of

We are interested in knowing **why you are studying engineering**. Please indicate below how much each of the following reasons applies to you in terms of why you are studying engineering:

	Not a reason (1)	Minimal reason (2)	Moderate reason (3)	Major reason (4)	Prefer not to say (5)
Technology plays an important role in solving society's problems	○	○	○	○	○
Engineers make more money than most other professionals	○	○	○	○	○
My parents would disapprove if I chose a degree other than engineering	○	○	○	○	○
Engineers have contributed greatly to fixing problems in the world	○	○	○	○	○
Engineers are well paid	○	○	○	○	○
My parents want me to be an engineer	○	○	○	○	○
An engineering degree will guarantee me a job when I graduate	○	○	○	○	○
A mentor has encouraged and/or inspired me to study engineering	○	○	○	○	○
A mentor has introduced me to people and opportunities in engineering	○	○	○	○	○
I feel good when I am doing engineering	○	○	○	○	○
I like to build stuff	○	○	○	○	○
I think engineering is fun	○	○	○	○	○
Engineering skills can be used for the good of society	○	○	○	○	○
I think engineering is interesting	○	○	○	○	○
I like to figure out how things work	○	○	○	○	○

Figure 2. APPLES original version with Likert-type response options.

15 statements and scored 1 for selected or 0 if not chosen. The six dimensions were treated as distinct, and items were analyzed separately in Winsteps to produce the item statistics. Rasch statistics examined include the Wright map, item reliability and differential item functioning (DIF). Cronbach's alpha was calculated for the sub-sample who answered the original version of the questionnaire (Likert-type). The reliability and validity of the motivation dimensions were examined via five aspects of measurement:

1. **Item fit to Rasch model:**
   Do items fit the Rasch model? The criteria used were infit and outfit mean square statistics (MSQ) to find items which significantly misfit the model, and if items had outfit or infit values higher than 1.5, they

were investigated (Combrinck, 2020; Linacre, 2002). The point-measure correlation (PTMA) was assessed for the Likert-type data, and a positive correlation is expected as items should correlate with one another and the underlying construct. Negative correlations would require further investigation as this could indicate items that should have been reversed scored or that are not behaving as expected (Linacre, 2023b). Some authors offer guidelines such as 0.30 to 0.70 as a range for the PTMA (Allen & Yen, 2002). Wright (1992) emphasized that eliminating items based on low correlations could remove necessarily easy or difficult items that form part of the construct. The current analysis only investigated the positive nature of correlations.

2. Reliability: *Do persons show consistency in their responses?* I partly assessed the reliability with the Wright Map, which has implications for reliability. The maps were investigated for the spread of items with persons and items less or more likely to be endorsed. Reliability was also investigated via item reliability and the separation index. Item reliability refers to how well the sample represents a widespread construct, and the item separation index shows evidence of construct validity in terms of item sequence (Linacre, 2023b). Item reliabilities above 0.70 are recommended, and good separation indices are above 3 (Fisher, 1992; Wolins et al., 1983).

3. Invariance: *Do items function in the same way for different groups?* Assessed with differential item functioning (DIF), if the magnitude of the difference is large (>0.5) and significant, then the differences between groups were investigated (Linacre, 2023b). The groups chosen for comparison were gender, ethnicity and first-generation to study engineering. The background variables were chosen because they are correlated with motivation in engineering studies (Dugard & Sánchez, 2021; Kalender et al., 2019; Lichtenstein et al., 2014; Nwanua Ohei & Brink, 2021)

4. Unidimensionality: *Do the items form a cohesive dimension?* Assessed with principal component analysis (PCA) of Rasch residuals for the original form of the instrument and compared with PCA of ipsative data. If the first contrast is smaller than 2, evidence for a lack of multidimensionality is accepted (Aryadoust et al., 2021). The PCA was conducted per construct for the six dimensions. The literature recommends not using Principal Component Analysis (PCA) for ipsative data (Brown & Maydeu-Olivares, 2018a, 2018b; Carvajal & Gomez, 2014). However, log transformations can be utilized for factor analysis with ipsative data (Batista-Foguet et al., 2015). An advantage of Rasch models is that they perform a natural log transformation on the data, ideal for ipsative data.

Due to the ipsative nature of the data, a further step was taken to check items' invariance across background factors. Non-parametric tests were conducted using an average number of times a category was chosen (e.g., the mean of *social good* options chosen). I added this additional analysis to check if the differential item functioning resulted in all items becoming more likely to be chosen (differential test functioning) in the forced-choice format. The combination of Rasch analysis and non-parametric statistics is due to the ipsative nature of the data. I include both as an additional check of whether the two formats can be compared and whether valid and reliable inferences can be derived from the forced-choice format. Ipsative data have certain constraints, one being that parametric statistics cannot be applied. Furthermore, some types of psychometric analysis cannot be conducted with this type of data, for example, reliability coefficients being calculated. As the onus lies here with the author to show the feasibility of using the forced-choice format, additional analysis was warranted for extra checks of instrument functionality.

The comparability for the two formats was first established by randomly assigning the sub-sample who answered the Likert format to reduce sampling bias. Secondly, common item equating was built into the process through the other APPLES items students completed. There were other items related to their motivation to study engineering, such as career aspirations after graduation. For example, measuring social good was done with the Likert-type items, the forced-choice items and the linking items from aspirations after studying engineering.

Ethical Statement

Both faculty and students provided permission for the study. All first-year students completed the survey, but I only used data in the current analysis when students gave informed consent for further research uses.

Results

In this section, I compare the forced choice and the Likert format using descriptive and inferential statistics. Then, I delve into the internal validity and reliability of the two types by applying the dichotomous and rating scale Rasch models.

Table 1. Count and Percentage of How Many Started and Finished the Survey and Missing Cases.

Response Type	Started	Finished	% Finished	Missing cases (%)	Minutes to complete
Version 1 (Forced-choice)	1,241	967	78	0.9	7.24
Version 2 (Likert-type)	191	121	63	10.7	13.80
Total	1,432	1,088			

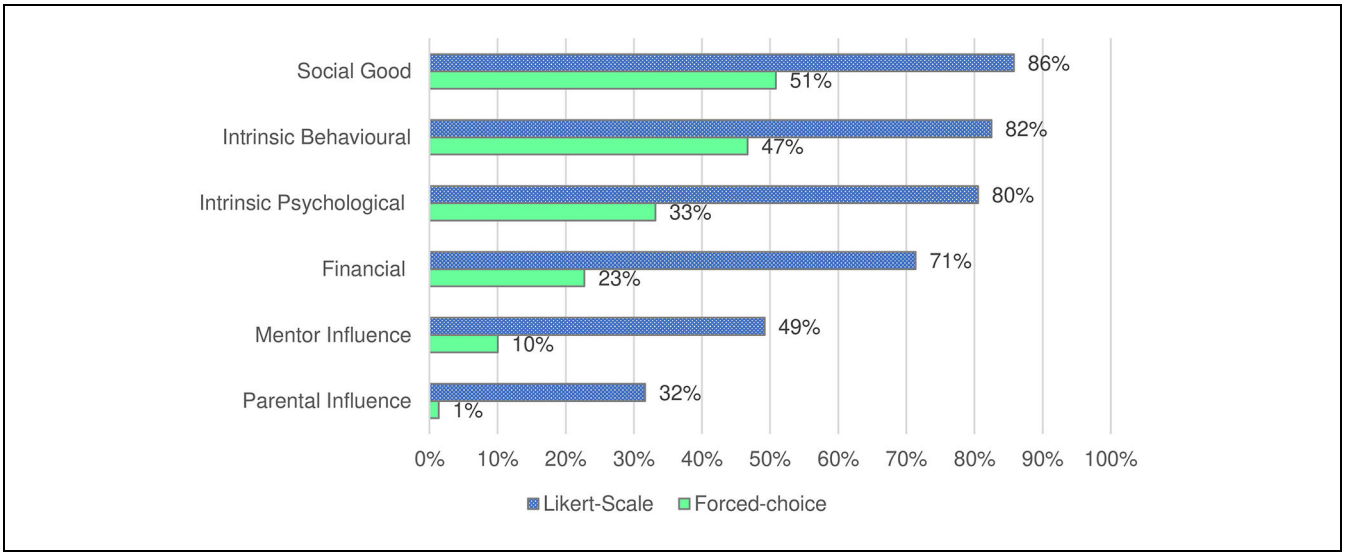

Figure 3. Degree of dimension endorsement comparison for Likert-scale and forced-choice.

Comparison of Forced-Choice and Likert Response Rates

Table 1 shows the percentage of students who started the survey and finished it to compare the forced-choice version to the original Likert data (excludes 13% of respondents who completed the survey but declined consent for data to be used). I also show how many cases overall were missing; in the forced-choice version, this would be students who did not select any option and, in the Likert scale, those who chose not to answer.

The original Likert options were given to a smaller, random subsample. The Likert options require more time and attention, and the completion rate (63%) is lower than that of the forced-choice format (78%). The missing data rates also differ, with more respondents not answering all questions in the Likert version (10.7%) compared to the forced-choice version, where 0.9% of respondents did not select at least one option from the list. The average time required to complete the forced-choice format was 7.25 min, whereas the Likert type required more time (13.80 min on average). General recommendations from the literature are that shorter surveys are preferable and that 10 to 20 min should be the maximum time respondents spend online answering (Barton et al., 2021; M. Revilla & Ochoa, 2017).

The difference between forced-choice and Likert-scale selection is shown in Figure 3, with a percentage calculated for the forced-choice (out of options chosen) and the Likert format, out of 4.

The two formats have a similar underlying pattern, with *social good* being the most highly chosen aspect and *parental influence* being the least. However, the Likert format has a higher endorsement for all motivations to study engineering. Asking respondents to choose their top five reasons leads to a less ambiguous indication of which aspects motivated them, whereas asking them to rate on a scale from 1 to 4 led to high endorsement for most options (mentor and parental being the exceptions). The difference is most notable for *financial motivation*, where in the Likert version, students endorsed this highly overall (71%), and there was considerably less choosing of financial options in the forced-choice version (23%).

Item Indices

The item reliability and separation statistics are shown in Table 2.

The high item reliability ($\alpha = .99$) for both formats indicates that the sample size is large enough to accurately locate items on the motivations construct. The

Table 2. Item Reliability and Separation Indices Per Questionnaire Type.

Measurement requirement	Forced-choice format	Likert-type format
Item reliability	$\alpha = .99$	$\alpha = .99$
Item separation	Index = 8.46	Index = 8.79

high item separation index shows that items highly endorsed are most likely the favored options, and items with less endorsement are most likely less favored; the items accurately measure the construct. The item reliability and separation index provide evidence for the construct validity of both formats.

Table 3 shows the mean, model standard error (S.E.), infit mean square (IN.MSQ), outfit mean square (OUT.MSQ) and point-measure correlation (PTMA) for both item types, calculated per dimension. The only item with a high outfit mean square (M.S.Q. = 1.55) was question 6 (*My parents want me to be an engineer*), which came from the forced-choice format. However, the outfit needed to be deemed larger to warrant the removal or revision of the item (Linacre, 2023b). As expected in an ipsative data structure, the point-measure correlations are minor for the forced-choice format. The PTMAs are positive and reasonable for the Likert-type version. Since ipsative data constrain individual choices, responses cluster at extreme ends. The clustering is advantageous as it gives us a clearer idea of valid preferences. However, it also limits the potential for correlations as the range within the variable is constrained (L. A. Miller & Lovler, 2020). The constant sum produced by ipsative data can also create negative correlations, which are inflated (McLean & Chissom, 1986; Ried, 2014). Therefore, the low correlations for the forced-choice version are not unexpected. Generally, the fit statistics are unremarkable, that is, satisfactory, for all items on both test formats.

Reliability—Consistency of Items Endorsed

Using the Likert-type data showed most of the aspects had sufficiently high person reliabilities (coefficient alphas):

Social good = .72
Financial = .71
Intrinsic = .81
Mentor = .80
Parental = .62

The parental dimension was the only construct with a lower-than-desirable reliability coefficient.

When examining item patterns for the forced-choice version, there was a consistent endorsement of questions related to the same aspects and item fit statistics, indicating agreement amongst students on the construct, as seen on the Wright maps Figures 4 and 5. The maps show that *intrinsic motivation* and *social good* were readily endorsed and consistently chosen options. Conversely, *mentor* and *parental influence* were the least favored. The findings from the Wright maps can be used as evidence of the reliability of the dimensions.

Invariance—Differential Item Functioning (DIF)

The Likert-type options revealed no differential item functioning between males and females, but large and

Table 3. Descriptive Item Statistics From the Rasch Model—Calculated Per Construct and Compiled.

		Forced-choice format					Likert-type format				
No. Item	Description	Mean	S.E.	IN.MSQ	OUT.MSQ	PTMA	Mean	S.E.	IN.MSQ	OUT.MSQ	PTMA
1	Technology solves society's problems	−1.52	0.07	0.98	0.97	0.27	−.45	0.12	0.94	0.96	0.41
2	Engineers make more money	0.74	0.10	1.01	1.06	0.07	0.20	0.10	1.01	1.11	0.19
3	My parents would disapprove if I chose another degree	3.64	0.38	1.00	1.35	0.00	2.48	0.18	1.02	1.07	0.31
4	Engineers contribute to fixing problems in the world	−1.58	0.07	1.00	1.00	0.23	−1.13	0.15	1.01	1.00	0.46
5	Engineers are well paid	−.47	0.07	1.01	1.03	0.15	−.01	0.10	0.84	0.83	0.37
6	My parents want me to be an engineer	2.62	0.23	1.01	1.55	−.02	2.19	0.16	1.29	1.20	0.33
7	An engineering degree will guarantee me a job	−.31	0.07	1.02	1.02	0.13	−.03	0.10	1.04	1.18	0.30
8	A mentor has encouraged me to study engineering	0.43	0.09	1.01	1.04	0.10	0.94	0.10	1.08	1.05	0.48
9	A mentor introduced me to engineering	1.66	0.15	0.99	0.90	0.10	1.04	0.10	1.19	1.12	0.46
10	I feel good when I am doing engineering	0.17	0.08	0.99	0.94	0.17	−.05	0.11	1.17	1.19	0.47
11	I like to build stuff	−.72	0.07	1.01	1.03	0.16	−.38	0.12	1.09	1.03	0.54
12	I think engineering is fun	−.38	0.07	0.99	0.97	0.20	−.59	0.12	0.77	0.74	0.58
13	Engineering skills can be used for the good of society	−1.09	0.07	0.99	0.99	0.23	−1.08	0.15	0.90	0.95	0.57
14	I think engineering is interesting	−1.48	0.07	1.02	1.02	0.20	−1.52	0.17	0.88	0.89	0.51
15	I like to figure out how things work	−1.71	0.07	0.99	0.99	0.26	−1.61	0.18	0.80	0.76	0.52

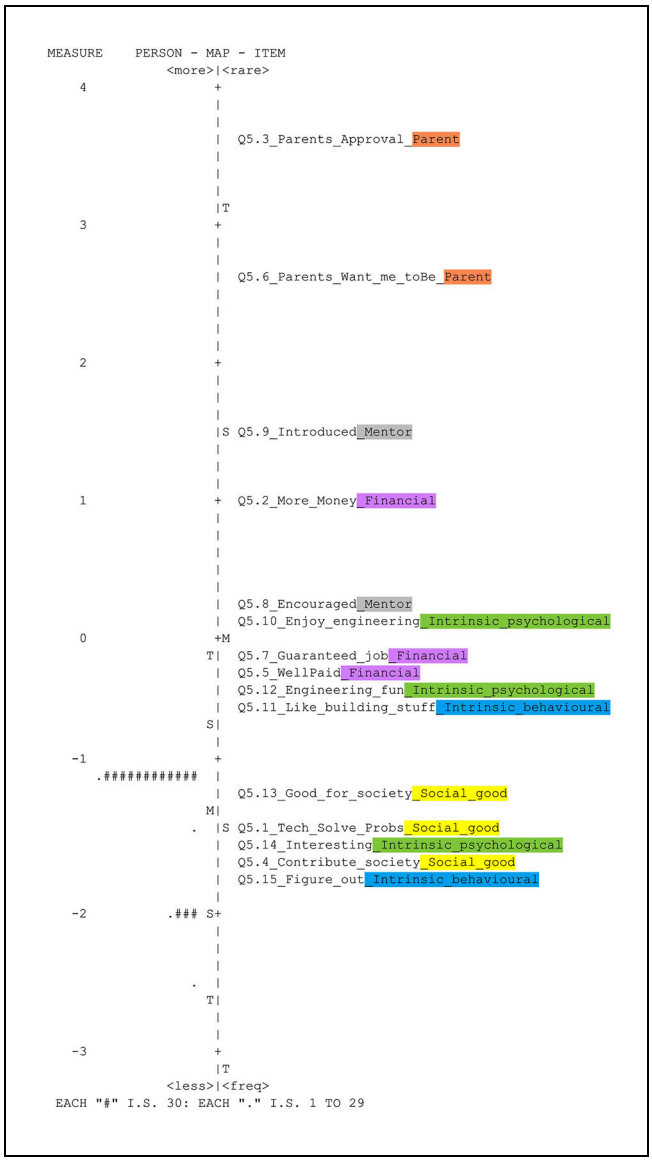

Figure 4. Wright map of forced-choice format.

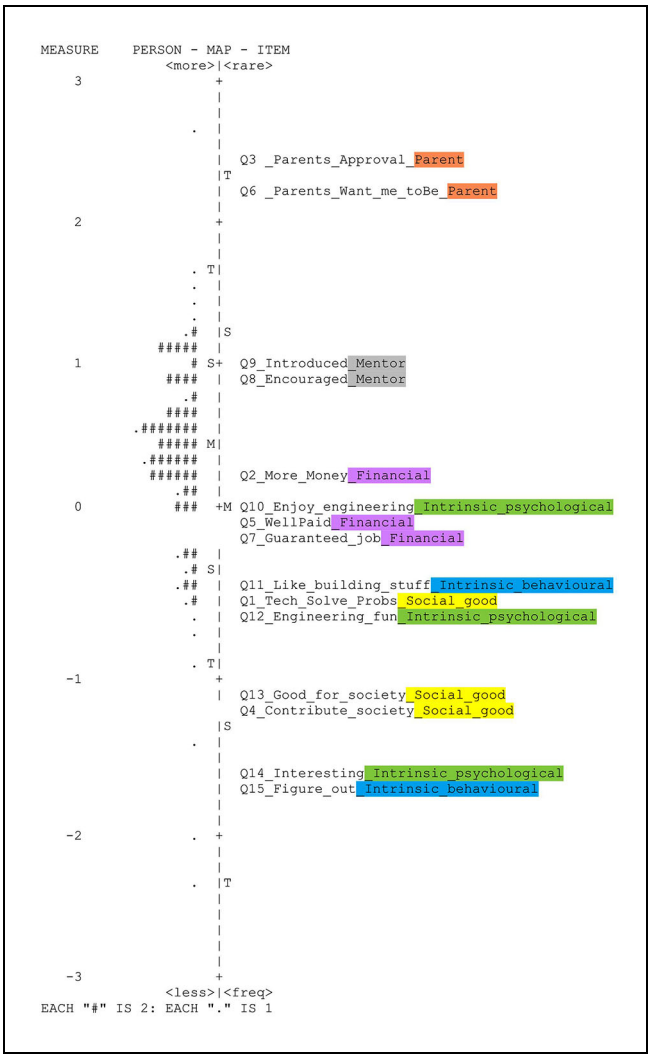

Figure 5. Wright map of Likert type format.

significant DIF was found in the forced-choice version shown in Table 4. Only items with large (contrast > 0.5) and significant ($p < .05$) DIF are reported.

Women had a large and significant differential contrast for item 7 (*Engineering will guarantee me a job*). At the same time, men were more likely to choose items related to intrinsic motivation (e.g., *engineering is fun*). The difference is significant across all intrinsic and financial motivation items, as shown in Table 5.

The Likert-type options did not show significant differences between the genders on the six motivational aspects. In the forced-choice format, women are significantly more likely to choose motivation options related to *finances* ($p = .00$), and men are significantly more likely to choose *intrinsic behavioral* ($p = .03$) statements as their reason for choosing to study engineering. However, the effect size is small for *financial reasons*, and for *intrinsic behavioral motivation*, the effect is negligible.

In terms of ethnic differences for the Likert types, I did not find large and significant DIF between the ethnic groups. The Kruskal–Wallis test also revealed no significant difference in the dimensions between groups for the Likert options. Four items were identified with large and significant DIF in the forced-choice format analysis. For the items where DIF was detected, the percentage per ethnic group that chose the motivational forced-choice option is displayed in Figure 6.

White students chose the options for *social good* and most *intrinsic behavioral* options more often than the other groups. For the intrinsic options, White students chose an intrinsic reason as their primary motivation for studying engineering more than 60% of the time.

Table 4. Differential Item Functioning—Items Showing Significant and Large Differences Between Genders.

Statements	% Selected option		Differential mean measure		Size and significance	
	Male	Female	Male	Female	DIF contrast	Mantel-Haenszel *p*-value
Q5.7 Financial: *Engineering will guarantee me a job*	18	38	0.20	−0.82	−1.02	.00
Q5.10 Intrinsic psych: *I feel good when I am doing engineering*	21	12	0.00	0.71	0.71	.01
Q5.11 Intrinsic behavioral: *I like to build stuff*	37	19	−0.83	0.12	0.95	.00
Q5.12 Intrinsic psychological: *I think engineering is fun*	31	22	−0.56	−0.03	0.52	.02

Table 5. Mann–Whitney *U* Test to Compare Gender Means on Motivation Constructs.

Construct	Mann-Whitney *U*	*Z*	Asymp. sig. (two-tailed)	Effect size *r*
Social good	30,019.00	−1.10	.27	−.05
Financial	26,619.50	−3.38	.00*	−.14
Parental	31,289.00	−1.18	.24	−.05
Mentor	30,081.00	−1.46	.14	−.06
Intrinsic psychological	28,980.00	−1.78	.08	−.08
Intrinsic behavioral	28,280.50	−2.22	.03*	−.09

*$p < .05$.

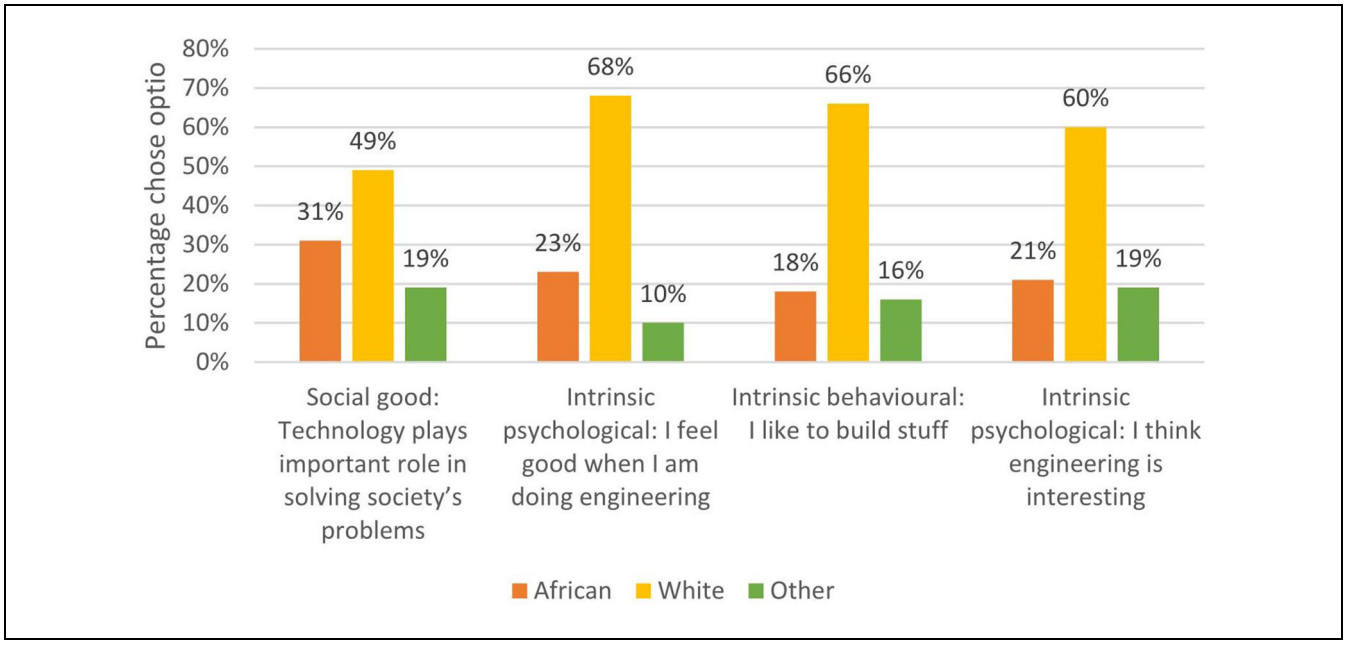

Figure 6. Percentage of ethnic groups who chose the option showing differential item functioning.

Table 6 compares differential item functioning (DIF) among ethnic groups responding to the APPLES survey, and only large and significant differences are reported.

Item 1 revealed significant and large DIF between African and White students and between White and Other ethnicities. A similar pattern can be observed for the other items, where White students were significantly more likely to endorse the intrinsic options. Table 7 shows the Kruskal-Wallis results and the effect sizes based on the Mann-Whitney comparisons.

African and White students were the only comparisons that yielded significant differences and small effect

Table 6. Differential Item Functioning—Items Showing Significant and Large Differences Between Ethnicities.

	African vs White			White vs Other			Other vs Black		
Items	Dif Mean	DIF size	*p*	Dif Mean	DIF size	*p*	Dif Mean	DIF size	*p*
Q1 Social good: *Technology plays an important role*	−2.16	−0.98	.00*	−1.18	0.73	.00*	−1.91	0.25	.53
Q10 Intrinsic psychological: *I feel good doing engineering*	0.30	0.31	.32	−0.01	−0.85	.03*	0.83	0.54	.26
Q11 Intrinsic behavioral: *I like to build stuff*	−0.08	0.69	.00*	−0.78	−0.19	.50	−0.58	−0.50	.16
Q14 Intrinsic psychological: *I think engineering is interesting*	−1.19	0.50	.04*	−1.68	0.23	.40	−1.91	−0.72	.02*

*$p < .05$.

Table 7. Kruskal-Wallis *H* Tests to Compare Ethnic Means on Motivation Constructs.

	Kruskal-Wallis *H*			Effect sizes *r* based on Mann-Whitney *U*		
Dimension	H	*df*	Asymp. sig.	African & White	African & Other	White & Other
Social good	8.60	2	.01*	−0.15	−0.04	−0.09
Financial	1.53	2	.47	−0.07	−0.03	−0.02
Parental	5.04	2	.08	−0.11	.00	−0.09
Mentor	3.84	2	.15	−0.11	−0.07	−0.03
Intrinsic psychological	6.74	2	.03*	−0.14	−0.09	−0.04
Intrinsic behavioral	9.32	2	.01*	−0.16	−0.05	−0.08

*$p < .05$.

sizes. The White students are significantly ($p = .01$) more likely to choose the *social good* options, but the effect size is small ($r = -.15$). The same findings show up for *intrinsic psychological* and *behavioral* motivation.

Another variable, first-generation students, was also investigated for DIF. None of the items showed large or significant differences for students who were first-generation attendees when compared with students who had parents or relatives who attended university for either of the survey versions.

Unidimensionality

A comparison using the current data shows that P.C.A. results for the traditional Likert items (version 2) and the forced-choice options yielded similar results, as shown in Table 8.

In both the ipsative and the Likert-type versions, there was sufficient evidence that the respective constructs are unidimensional, and the unexplained variance is below the threshold of 2, as Linacre (2023b) recommended. In a Rasch PCA, the main component is the underlying construct, and the contrasts are potential dimensions which could compete with the main Rasch-explained variance. In the current comparison, the most substantial contrast does not reach the threshold to compete with the central dimension, except if a PCA is conducted on all the Likert-type items and the dimensions are not considered. The analysis presented here shows that Rasch PCA can be used for ipsative data if the dimensions are pre-defined; the PCA for a forced-choice format does not reveal underlying dimensions.

Discussion

The study found that forced-choice items had improved response and completion rates. The forced-choice format could lead to a more efficient survey with enhanced respondent engagement. The fact that missingness was reduced in the forced-choice format indicates a potential contribution to improved data quality. Construct endorsement emerged more clearly in the forced-choice format. The same patterns of construct endorsement were observed in the two format types, suggesting that both types can help identify the underlying constructs on a range. However, specific motivations emerged more strongly in the forced-choice format, which could indicate a more accurate measurement, or it could be a by-product of the forced nature of the item. The reliability and separation indices were acceptable for both formats. When I looked at patterns in the data between the two versions, there were similarities in which motivations were highly favored. The consistency is encouraging as the forced-choice version picks up the same pattern but

Table 8. Comparison of P.C.A. Unexplained Variance of Two Types of Variables.

Construct	V1 forced-choice P.C.A.—Unexplained variance in the first contrast	V2—Likert-type P.C.A.—Unexplained variance in the first contrast
Social good	1.5230	1.8217
Financial	1.6419	1.7317
Parents	0.0000	0.0011
Mentors	0.0000	0.0009
Intrinsic	1.4452	1.5111
All items (15)	1.8382	3.4248

more clearly. For example, most participants selected *social good* as their reason for studying engineering, and only a few selected *parental influence* in both questionnaire versions. The most substantial difference can be seen when choosing financial reasons for studying engineering; while most students (71%) said it is a moderate or major reason in the Likert-version, only 23% chose a financial option as part of their top five in the forced-choice format. Using Likert-type formats can lead to a high agreement with statements due to time constraints, a lack of engagement from the respondent, or social desirability (Subedi, 2016). Using forced-choice items can reduce acquiescence bias and provide a clearer picture of the construct; similar findings were reported by Kreitchmann et al. (2019), Watrin et al. (2019), and Geldhof et al. (2015).

Both formats had high item reliability indices, indicating the satisfactory accuracy of measuring the underlying constructs. The Rasch model was valuable in analyzing both types of item formats. However, the user should be cautioned that the forced-choice format has more limitations regarding statistical models that can be applied. Differential item Function (DIF) was more clearly detected in the forced-choice format, whereas no DIF was detected in the Likert-type data. The presence of DIF in one but not the other could indicate that the forced-choice format reveals patterns which are not evident in the other format or that the nature of the format causes more differences between groups due to its obligatory nature.

The Rasch measurement model was a valuable tool for assessing the functioning of the forced-choice version and comparing psychometric properties to the Likert-type version (Bailes & Nandakumar, 2020). The forced-choice version data fit the Rasch model but not the Likert type. Both had high item reliability values (0.99) and large separation indices, which showed consistency in the options selected. Juxtaposing the two Wright maps, it emerged that the overall pattern is consistent in that *social good* and *intrinsic* reasons are easy to endorse and *parental and mentoring* are the least likely to be chosen. However, the item ordering on the maps differs; for example, financial items were more challenging to endorse on the Forced-choice Wright map. This difference in ordering may be a downside of the forced-choice data, as ipsative data produces more significant variance.

Cronbach's alpha cannot be calculated for ipsative data, and other Rasch statistics, such as the item separation index, were used to indicate consistency in responses. When I investigated the coefficient alpha values for the Likert version, the dimensions had acceptably high values (above .700) except for parental motivation, showing that the forced-choice version mirrored the reliability through the item separation index and the Wright map. The person separation index cannot be calculated for the forced-choice type, but I suggest using the Wright map to guide the consistency of items and dimensions endorsed and examining the item reliability produced by software such as Winsteps. The forced-choice type gave a good description of the items and dimensions.

The item fit statistics for the forced-choice version only had one item misfit, for the statement: *My parents want me to be an engineer*. A common phenomenon in surveys is that participants are less likely to endorse certain items. This trend is likely to be duplicated in forced-choice formats. Certain vital items may be under-selected, which happened in the current study. Therefore, it is crucial to understand why the respondents might view the options as irrelevant, uncomfortable, or undesirable. Consultation with engineering educators and other researchers raised the possibility that the negative slant of the questions may discourage respondents from choosing parental options. For example, if the item was rephrased as *My parents would be proud if I became an engineer*, it might draw a more positive and realistic endorsement of belief (area for future research). Negatively phrased statements and items may produce item bias and problems (Franchignoni et al., 2010; Pey Tee & Subramaniam, 2018). Therefore, we recommend that future versions of the APPLES adapt the parental influence items to reflect positive contributions parents can make to the motivation to study engineering.

When assessing the unidimensionality of the ipsative data, the Rasch model was identified as a viable method for principal component analysis (P.C.A.) due to the natural log transformations. A juxtaposition of Likert and forced choice P.C.A. results showed similar findings, where the first contrasts were too small (<2) to compete with the central dimension. The P.C.A. results show that Rasch models can assess the unidimensionality of constructs for ipsative data and support the alternative format used if the dimensions are specified prior to analysis.

In psychology, the forced-choice format has a well-known history of use with either positive results or comparative findings. For example, Guenole et al. (2018) found that ipsative formats worked well when evaluating maladaptive personality traits in the workplace. K. B. Boone (2021) found that forced-choice formats work well with neuropsychological exams. Bäckström and Björklund (2024) showed that the two formats can yield comparable information for personality research, leading them to advise using a mixture of the two types of items. Morillo et al. (2019) found that the F.C. format worked well for assessing personality traits, but the format did not wholly cover some aspects of the constructs. Wetzel et al. (2020) advise that more research and consideration are needed before changing the format of the response structure. Brown and Maydeu-Olivares (2018a, 2018b) found that the format works well for sensitive topics in educational settings. Zhang et al. (2024) did a comprehensive study of graded forced-choice items and their applicability in educational research and found that graded formats outperformed the traditional dichotomous style. Conversion of the format into conjoint analysis is also possible, and Hainmueller et al. (2014) demonstrate how to do this so that causal inferences can be derived in political science (Hainmueller et al., 2015). Ross and Bibler Zaidi (2019) make the crucial point that any survey format has limitations, and it is far more critical that the researchers are aware of the limitations and explicitly acknowledge these extensively in their reporting.

The current study's results should be used with other factors, and researchers are encouraged to find the balance between practical administration, high-quality data, and ease of analysis. Likert-type and forced-choice items have pros and cons, and the trade-offs should be considered before choosing a format. Of course, piloting both types could be optimal if a large enough sample is available.

Limitations of the Study

The nuances and limitations of question formats must be acknowledged. The Likert-type format may suit some survey types, such as questions requiring ratings rather than ranking or paired comparisons (J. D. Miller et al., 2018; Sung & Wu, 2018). Alternatively, both formats could yield equally reliable and valid inferences for some surveys, in which case the decision should be based on the practicalities of administration (Cohen et al., 2017; Ried, 2014). The Likert-type format has produced better comparative data for researchers who want to investigate the association between groups, another factor that should be considered (Heo et al., 2022). The advantages of Likert-type formats are that they offer a continuum of responses, produce ordinal data, offer the opportunity to investigate equal intervals, and are more flexible in currently available statistical methods (Bäckström & Björklund, 2024). The distinct advantages of Likert-type scales' rich data analytical flexibility should be considered against their disadvantages when researchers decide whether to try forced-choice. The choice of first-year engineering students as a sample is a limitation as first-year students are more eager to please and may be more prone to social desirability. Further research studies can be done where other samples and constructs are trialed with forced-choice formats and compared to Likert-type scales for usability.

Guidelines for Researchers

The results presented a psychometrics-based case for the validity and reliability of using forced-choice questions in self-reported instruments and how psychometric models can be applied to evaluate ipsative data. A valid and reliable construct representation requires a range to be measured, which can be more challenging to obtain with the Likert format if social desirability or acquiescence bias is present. Forced-choice options indicate a construct's range more clearly and negate some social desirability responses. The Rasch model can assess the reliability and validity of ipsative data.

Guidelines for researchers on how to analyze and defend forced-choice formats:

1. **The researcher should treat questions as options when using the Rasch model for ipsative data.** The forced-choice nature of the data results in items becoming options and the current paper can be used as an example of how to deal with this during the analysis.
2. **Look for consistency in choices as an alternative to person reliability (coefficient alpha) alternative.** The consistency and ordering of items, as observed in the Wright map, is one way to accomplish this, and another is to examine the item's reliability. Having a sub-sample complete the Likert-type data is another potential route, as it provides additional information and opportunities for comparison.

3. **Investigate differential item functioning (DIF)**, but keep in mind the forced nature of choices; look at the big picture in addition to item investigation. Examine the overall dimensions with the aid of non-parametric comparisons.

The current study demonstrated how the Rasch models could be used to check the functioning of forced-choice questionnaires, and awareness and application of the format could lead to more valuable inferences and enhance meaningful measurement for future researchers.

Acknowledgments

Professor Trevor Bond and Professor William P. Fisher, Jr. read and gave valuable feedback. I am very grateful to both for their time, inputs and congeniality.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The South African Department of Higher Education and Training (DHET) provided funding for writing this article and academic support through the Scholarship of Teaching and Learning (SoTL) grant.

Ethics Statement

Ethical clearance was obtained from the University of Pretoria's Built Environment and Information Technology Faculty's (EBIT) ethics committee to conduct the study, reference number: EBIT/46/2020.

ORCID iD

Celeste Combrinck https://orcid.org/0000-0002-8067-5299

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Andrich, D. (1989). A probabilistic I.R.T. model for unfolding preference data. *Applied Psychological Measurement*, *13*(2), 193–216.

Andrich, D. (2016). Rasch rating-scale model. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory* (pp. 75–94). Chapman and Hall/C.R.C.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Bäckström, M., & Björklund, F. (2024). Why forced-choice and Likert items provide the same information on personality, including social desirability. *Educational and Psychological Measurement*, *84*(3), 549–576. https://doi.org/10.1177/00131644231178721

Bailes, L. P., & Nandakumar, R. (2020). Get the most from your survey: an application of Rasch analysis for education leaders. *International Journal of Education Policy and Leadership*, *16*(2), 1–19. https://doi.org/10.22230/ijepl.2020v16n2a857

Barton, B. A., Adams, K. S., Browne, B. L., & Arrastia-Chisholm, M. C. (2021). The effects of social media usage on attention, motivation, and academic performance. *Active Learning in Higher Education*, *22*(1), 11–22. https://doi.org/10.1177/1469787418782817

Batista-Foguet, J. M., Ferrer-Rosell, B., Serlavós, R., Coenders, G., & Boyatzis, R. E. (2015). An alternative approach to analyze ipsative data. Revisiting experiential learning theory. *Frontiers in Psychology*, *6*, 1742. https://doi.org/10.3389/fpsyg.2015.01742

Berry, K., Rana, R., Lockwood, A., Fletcher, L., & Pratt, D. (2019). Factors associated with inattentive responding in online survey research. *Personality and Individual Differences*, *149*, 157–159. https://doi.org/10.1016/j.paid.2019.05.043

Bond, T. G., Yan, Z., & Heene, M. (Eds.). (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Boone, H., & Boone, D. (2012). Analyzing likert data. *Journal of Extension*, *50*(2), 48. https://doi.org/10.34068/joe.50.02.48

Boone, K. B. (2021). *Assessment of feigned cognitive impairment*. Guilford Publications.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160. https://doi.org/10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2013). How I.R.T. can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. https://doi.org/10.1037/a0030641

Brown, A., & Maydeu-Olivares, A. (2018a). Modelling forced-choice response formats. In P. K. J. Han, F. M. P. De Caro, & M. A. H. J. Schmitt (Eds.), *The Wiley handbook of psychometric testing* (pp. 523–569). Wiley-Blackwell.

Brown, A., & Maydeu-Olivares, A. (2018b). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling A Multidisciplinary Journal*, *25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Buchanan, B. S., & Morrison, D. G. (1985). Measuring simple preferences: An approach to blind, forced choice product testing. *Marketing Science*, *4*(2), 93–109.

Calderón Carvajal, C., Ximénez Gómez, C., Lay-Lisboa, S., & Briceño, M. (2021). Reviewing the structure of Kolb's learning style inventory from factor analysis and thurstonian item response theory (I.R.T.) model approaches. *Journal of Psychoeducational Assessment*, *39*(5), 593–609. https://doi.org/10.1177/07342829211003739

Carvajal, C. C., & Gomez, C. X. Y. (2014). Factor analysis of forced-choice items: A review and an example. *Revista Latinoamericana de Psicologia*, *46*(1), 24–34. https://doi.org/10.1016/s0120-0534(14)70003-2

Castaño, A. M., Lubiano, M. A., & García-Izquierdo, A. L. (2020). Gendered beliefs in STEM undergraduates: A comparative analysis of fuzzy rating versus Likert scales. *Sustainability*, *12*(15), 6227. https://doi.org/10.3390/su12156227

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika*, *30*(1), 99–121. https://doi.org/10.2333/bhmk.30.99

Chen, W.-H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of life research*, *23*, 485–493. https://doi.org/10.1007/s11136-013-0487-5

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling A Multidisciplinary Journal*, *9*(1), 55–77. https://doi.org/10.1207/s15328007sem0901_4

Cohen, L., Manion, L., & Morrison, K. (2017). Questionnaires. In L. Cohen, L. Manion & K. Morrison (Eds.), *Research methods in education* (8th ed., pp. 471–505). Routledge.

Combrinck, C. (2020). Is this a useful instrument? An introduction to Rasch models for evaluating tests and questionnaires. In S. Kramer, S. Laher, A. Fynn, & H. H. Janse van Vuuren (Eds.), *Online readings in research methods (ORIM)* (Vol. 1, pp.127–181). Psychological Society of South Africa.

Combrinck, C., & Inglis, H. (2020). *The validity of international instruments for assessing South African engineering students* [Conference session]. 2020 IFEES World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC), Cape Town. https://ieeexplore.ieee.org/document/9293636

Direito, I., Chance, S. M., Tilley, E., & Mitchell, J. E. (2019). *Assessing the grit and mindset of incoming engineering students with an emphasis on gender* [Conference session]. Research in Engineering Education Symposium (REES), Cape Town, South Africa.

Dolnicar, S., Grun, B., Leisch, F., & Rossiter, J. (2011). Three good reasons NOT to use five and seven point Likert items. *Research Online*. https://ro.uow.edu.au/commpapers/775

Donaldson, K. M., Chen, H. L., Clark, M., Toye, G., & Sheppard, S. D. (2008). *Scaling up: Taking the academic pathways of people learning engineering survey (APPLES)* [Conference session]. National 2008 IEEE Frontiers in Education Conference. https://doi.ieeecomputersociety.org/10.1109/FIE.2008.4720596

Douglas, J. W., & Tramonte, L. (2015). Towards the development of contextual questionnaires for the PISA for development study. https://doi.org/doi:https://doi.org/10.1787/5js1kv8crsjf-en

Dugard, J., & Sánchez, A. M. (2021). Bringing gender and class into the frame: An intersectional analysis of the decoloniality-as-race critique of the use of law for social change. *Stellenbosch Law Review*, *32*(1), 24–46. https://doi.org/10.47348/slr/v32/i1a2

Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Eris, O., Chachra, D., Chen, H. L., Sheppard, S., Ludlow, L., Rosca, C., Bailey, T., & Toye, G. (2010). Outcomes of a longitudinal administration of the persistence in engineering survey. *Journal of Engineering Education*, *99*(4), 371–395. https://doi.org/10.1002/j.2168-9830.2010.tb01069.x

Fisher, W. P. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*(3), 238. https://www.rasch.org/rmt/rmt63i.htm

Fisher, W. P. (2009). Invariance and traceability for measures of human, social, and natural capital: Theory and application. *Measurement*, *42*(9), 1278–1287.

Fisher, W. P. (2022). Contrasting roles of measurement knowledge systems in confounding or creating sustainable change. *ACTA IMEKO*, *11*(4), 1–6.

Franchignoni, F., Giordano, A., Michail, X., & Christodoulou, N. (2010). Practical lessons learned from use of Rasch analysis in the assessment of outcome measures. *Revista da Sociedade Portuguesa de Medicina Física e de Reabilitação*, *19*(2), 5–12. https://doi.org/10.25759/spmfr.39

Geldhof, G. J., Gestsdottir, S., Stefansson, K., Johnson, S. K., Bowers, E. P., & Lerner, R. M. (2015). Selection, optimization, and compensation: The structure, reliability, and validity of forced-choice versus Likert-type measures in a sample of late adolescents. *International Journal of Behavioral Development*, *39*(2), 171–185. https://doi.org/10.1177/0165025414560447

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment*, *25*(4), 513–526. https://doi.org/10.1177/1073191116641181

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395–2400. https://doi.org/10.1073/pnas.1416587112

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Hall, L., Hume, C., & Tazzyman, S. (2016). *Five degrees of happiness: Effective smiley face likert scales for evaluating with*

children [Conference session]. *Proceedings of the 15th International Conference on Interaction Design and Children*.

Hancock, P. A., & Volante, W. G. (2020). Quantifying the qualities of language. *15*(5), e0232198. https://doi.org/10.1371/journal.pone.0232198

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850. https://doi.org/10.1016/j.cptl.2015.08.001

Heo, C. Y., Kim, B., Park, K., & Back, R. M. (2022). A comparison of best-worst scaling and Likert scale methods on peer-to-peer accommodation attributes. *Journal of Business Research - Turk*, *148*, 368–377. https://doi.org/10.1016/j.jbusres.2022.04.064

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and I.R.T. scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612. https://doi.org/10.1177/01466216155858

Huang, J. L., & Wang, Z. (2021). Careless responding and insufficient effort responding. In *Oxford Research Encyclopedia of Business and Management*. https://doi.org/https://doi.org/10.1093/acrefore/9780190224851.013.303

IBM. (2023). *IBM SPSS Statistics for Windows (Version 28.0)*. I.B.M. Corp.

Inglis, H., Combrinck, C., & Simpson, Z. (2022). Disrupted access and success: Students' transition to university in the time of Covid-19. *SOTL in the South 6*(2): 53–72. https://doi.org/10.36615/sotls.v6i2.227

Inglis, H., & Simpson, Z. (2023, 11–13 July). *"You don't know anything until you know everything": Threshold concepts in first year student narratives about engineering* [Conference session]. Conference of the South African Society for Engineering Education, Muldersdrift.

Jaeger, S. R., & Cardello, A. V. (2022). Factors affecting data quality of online questionnaires: Issues and metrics for sensory and consumer research. *Food Quality and Preference*, *102*, 104676. https://doi.org/10.1016/j.foodqual.2022.104676

Kalender, Z. Y., Marshman, E., Schunn, C. D., Nokes-Malach, T. J., & Singh, C. (2019). Gendered patterns in the construction of physics identity from motivational factors. *Physical Review Physics Education Research*, *15*(2), 020119. https://doi.org/10.1103/physrevphyseducres.15.020119

Kemper, N. S., Campbell, D. S., Earleywine, M., & Newheiser, A.-K. (2020). Likert, slider, or text? Reassurances about response format effects. *Addiction Research & Theory*, *28*(5), 406–414. https://doi.org/10.1080/16066359.2019.1676892

Knoch, U., & McNamara, T. (2015). Rasch analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 275–304). Routledge.

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. Psychometric modeling of Likert items. *Frontiers in Psychology*, *10*, 2309. https://doi.org/10.3389/fpsyg.2019.02309

Kusmaryono, I., Wijayanti, D., & Maharani, H. R. (2022). Number of response options, reliability, validity, and potential bias in the use of the Likert scale education and social science research: A literature review. *International Journal of Educational Methodology*, *8*(4), 625–637. https://doi.org/10.12973/ijem.8.4.625

Latkin, C. A., Edwards, C., Davey-Rothwell, M. A., & Tobin, K. E. (2017). The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. *Addictive Behaviors*, *73*, 133–136. https://doi.org/10.1016/j.addbeh.2017.05.005

Lee, P., Joo, S.-H., & Lee, S. (2019). Examining stability of personality profile solutions between Likert-type and multidimensional forced choice measure. *Personality and Individual Differences*, *142*, 13–20. https://doi.org/10.1016/j.paid.2019.01.022

Lichtenstein, G., Chen, H. L., Smith, K. A., & Maldonado, T. A. (2014). Retention and persistence of women and minorities along the engineering pathway in the United States. In A. Johri & B. M. Olds (Eds.), *Cambridge handbook of engineering education research* (pp. 311–334).

Likert, R. (1974). The method of constructing an attitude scale. In R. Likert (Ed.), *Scaling* (pp. 233–242). Routledge.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. M. (2023a). *Winsteps® (Version 5.4.0.0)*. https://www.winsteps.com/

Linacre, J. M. (2023b). *Winsteps® Rasch measurement computer program User's Guide*. Winsteps.com. https://www.winsteps.com/

McLean, J. E., & Chissom, B. S. (1986). Multivariate Analysis of Ipsative Data: Problems and Solutions. *Annual Meeting of the Mid-South Educational Research Association* (1), 18-21. https://files.eric.ed.gov/fulltext/ED278717.pdf

Miller, J. D., Gentile, B., Carter, N. T., Crowe, M., Hoffman, B. J., & Campbell, W. K. (2018). A comparison of the nomological networks associated with forced-choice and Likert formats of the Narcissistic personality inventory. *Journal of Personality Assessment*, *100*(3), 259–267. https://doi.org/10.1080/00223891.2017.1310731

Miller, L. A., & Lovler, R. L. (Eds.). (2020). *Foundations of psychological testing : A practical approach* (6th ed.). Sage. https://search.ebscohost.com/login.aspx?direct = true&scope = site&db = nlebk&db = nlabk&AN = 3361639

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Revista de Psicología del Trabajo y de las Organizaciones*, *35*(2), 75–83. https://doi.org/10.5093/jwop2019a11

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, *142*(2), 71–89.

Nemoto, T., & Beglar, D. (2014, November 11–14). Likert-scale questionnaires. JALT 2013 conference proceedings, Japan.

Nwanua Ohei, K., & Brink, R. (2021). Trends in gender and behavioural disparities among South African university students: Choosing an Ict-related career path. *African Journal*

of Development Studies, *SI*(1), 111–141. https://doi.org/10.31920/2634-3649/2021/siv1a6

Pey Tee, O., & Subramaniam, R. (2018). Comparative study of middle school students' attitudes towards science: Rasch analysis of entire TIMSS 2011 attitudinal data for England, Singapore and the U.S.A. as well as psychometric properties of attitudes scale. *International Journal of Science Education*, *40*(3), 268–290. https://doi.org/10.1080/09500693.2017.1413717

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Paedagogiske Institute.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. ERIC.

Retief, L., Potgieter, M., & Lutz, M. (2013). The usefulness of the rasch model for the refinement of likert scale questionnaires. *African Journal of Research in Mathematics Science and Technology Education*, *17*(12), 126–138. https://doi.org/10.1080/10288457.2013.828407

Revilla, M., & Ochoa, C. (2017). Ideal and maximum length for a web survey. *International Journal of Market Research*, *59*(5), 557–565. https://doi.org/10.2501/ijmr-2017-039

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, *43*(1), 73–97. https://doi.org/10.1177/0049124113509605

Ried, L. D. (2014). Using Likert-type and ipsative/forced choice items in sequence to generate a preference. *Research in Social and Administrative Pharmacy*, *10*(4), 598–607. https://doi.org/10.1016/j.sapharm.2013.09.001

Ross, P. T., & Bibler Zaidi, N. L. (2019). Limited by our limitations. *Perspectives on Medical Education*, *8*(4), 261–264. https://doi.org/10.1007/s40037-019-00530-x

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 25–42). Springer.

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology*, *88*(4), 797–834. https://doi.org/10.1111/joop.12098

Schulte, N., Holling, H., & Bürkner, P.-C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement*, *81*(2), 262–289. https://doi.org/10.1177/0013164420934861

Sheppard, S. D., Gilmartin, S. K., Chen, H. L., Donaldson, K., Lichtenstein, G., Eris, O., Lande, M., & Toye, G. (2010). *Exploring the Engineering Student Experience: Findings from the Academic Pathways of People Learning Engineering Survey (APPLES)*. https://files.eric.ed.gov/fulltext/ED540124.pdf

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, *70*(1), 66–77. https://doi.org/10.1093/poq/nfj007

South, L., Saffo, D., Vitek, O., Dunne, C., & Borkin, M. A. (2022). Effective use of Likert scales in visualization evaluations: A systematic review. *Computer Graphics Forum,y*, *41*(3), 43–55. https://doi.org/10.1111/cgf.14521

Subedi, B. P. (2016). Using Likert type data in social science research: Confusion, issues and challenges. *International journal of contemporary applied sciences*, *3*(2), 36–49.

Sullivan, G. M., & Artino, A. R., Jr. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Sung, Y.-T., & Wu, J.-S. (2018). The visual analogue scale for rating, ranking and paired-comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods*, *50*(4), 1694–1715. https://doi.org/10.3758/s13428-018-1041-8

van Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing*, *20*(1), 196–201. https://doi.org/10.1046/j.1365-2648.1994.20010196.x

van Eijnatten, F. M., van der Ark, L. A., & Holloway, S. S. (2015). Ipsative measurement and the analysis of organizational values: An alternative approach for data analysis. *Quality & Quantity*, *49*(2), 559–579. https://doi.org/10.1007/s11135-014-0009-8

Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, *58*(8), 922–933. https://doi.org/10.1007/s11162-017-9448-0

Vesely, S., & Klöckner, C. A. (2020). Social desirability in environmental psychology research: Three meta-analyses. *Frontiers in Psychology*, *11*(1935), 1–9. https://doi.org/10.3389/fpsyg.2020.01395

Vidotto, G., Anselmi, P., Filipponi, L., Tommasi, M., & Saggino, A. (2018). Using overt and covert items in self-report personality tests: Susceptibility to faking and identifiability of possible fakers. *Frontiers in Psychology*, *9*, 1100. https://doi.org/10.3389/fpsyg.2018.01100

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*(8), 600–613. https://doi.org/10.1177/0146621617703183

Warmbrod, J. R. (2014). Reporting and interpreting scores derived from likert-type scales. *Journal of Agricultural Education*, *55*(5), 30–47. https://doi.org/http://files.eric.ed.gov/fulltext/EJ1122774.pdf

Watrin, L., Geiger, M., Spengler, M., & Wilhelm, O. (2019). Forced-choice versus Likert responses on an occupational big five questionnaire. *Journal of Individual Differences*,y, 134–148. https://doi.org/10.1027/1614-0001/a000285

Wetzel, E., Frick, S., & Greiff, S. (2020). *The multidimensional forced-choice format as an alternative for rating scales*. Hogrefe Publishing.

Wolins, L., Wright, B. D., & Masters, G. N. (1983). Rating scale analysis: Rasch measurement. *Journal of the American Statistical Association*, *78*(382), 497. https://doi.org/10.2307/2288670

Wright, B. D. (1992). Point-biserial correlations and item fits. *Rasch Measurement Transactions*, *5*(4), 174.

Wu, H., & Leung, S.-O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of* S*ocial* S*ervice* R*esearch*, *43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

Xiao, Y., Liu, H., & Li, H. (2017). Integration of the forced-choice questionnaire and the Likert scale: A simulation study. *Frontiers in Psychology*, *8*, 806. https://doi.org/10.3389/fpsyg.2017.00806

Yannakakis, G. N., & Martínez, H. P. (2015). Ratings are Overrated! [Mini Review]. *Frontiers in ICT*, 2. https://doi.org/10.3389/fict.2015.00013

Zhang, B., Luo, J., & Li, J. (2024). Moving beyond Likert and traditional forced-choice scales: A comprehensive investigation of the graded forced-choice format. *Multivariate Behavioral Research*,y, 434–460. https://doi.org/10.1080/00273171.2023.2235682

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569–590. https://doi.org/10.1177/1094428119836486