International Journal of Educational Methodology
Volume 11, Issue 3, 423 - 442.

ISSN: 2469-9632
https://www.ijem.com/

A Proposed Standard for the Reporting of Structural Equation Models With Ordinal Variables: Why Ordinal Data Should be Treated With Extra Care?

Gabriel Chun-Yeung Lee[*]
University of Nottingham, UNITED KINGDOM

Received: February 11, 2025 ▪ Revised: May 8, 2025 ▪ Accepted: July 11, 2025

Abstract: Educational researchers, as well as researchers in other disciplines, often work with ordinal data, such as Likert item responses and test item scores. Critical questions arise when researchers attempt to implement statistical models to analyse ordinal data, given that many statistical techniques assume the data analysed to be continuous. Could ordinal data be treated as continuous data, that is, assuming the ordinal data to be continuous and then applying statistical techniques as if analysing continuous data? Why and why not? Focusing on structural equation models (SEMs), particularly confirmatory factor analysis (CFA), this article discusses an ongoing debate on the treatment of ordinal data and reports a short review on the practices of conducting and reporting SEMs, in the context of mathematics education research. The author reviewed 70 publications in mathematics education research that reported a study involving SEMs to analyse ordinal data, but less than half discussed how data were treated or guided readers through the analysis; it is therefore harder to repeat such an analysis and evaluate the results. This article invites methodological discussions on SEMs with ordinal variables in the practices of educational research. Subsequently, a standard for reporting SEMs with ordinal data is proposed, followed by an example. This standard contributes to educational research by enabling researchers (self and others) to evaluate SEMs reported. The example demonstrates, using real-life research data, how two different approaches for analysing ordinal data (as continuous or as a product of discretisation from some continuous distributions) can lead to results that disagree.

Keywords: *Confirmatory factor analysis, Likert items, ordinal data, structural equation modelling.*

To cite this article: Lee, G. C.-Y. (2025). A proposed standard for the reporting of structural equation models with ordinal variables: Why ordinal data should be treated with extra care? *International Journal of Educational Methodology*, *11*(3), 423-442. https://doi.org/10.12973/ijem.11.3.423

Introduction

Educational researchers often work with observations (e.g., surveys, tests) related to unobserved characteristics (e.g., anxiety, beliefs). Those observations are often measured as ordinal data. Ordinal data indicate ordered categories, such as ‘always’ to ‘never’. For example, Johnny asks his participants to rate 15 items with statements related to their attitudes towards mathematics; each item has five options: ‘strongly agree’, ‘agree’, ‘neutral’, ‘disagree’, and ‘strongly disagree’. He thinks that the items can be grouped into confidence, enjoyment, and motivation. To test this hypothesis, he records his participants’ responses as 0–4 and fits the numeric data into a statistical model. In which the items are treated as individual variables and are grouped into three clusters. He uses available computing software with default options to assess his model, and then writes a report of his analysis, including item means and variances. How valid and reliable is Johnny’s analysis? What questions arise from the analysis steps? What does an item's mean score of 3.5 mean? Later, Johnny realises that the software has an option for ordered-categorical variables, so he chooses this option and re-runs the analysis. This time, the results show that the items can still be grouped into three clusters, but the grouping is different. However, a warning message appears in the software outputs – how does this message impact the validity of the analysis? How should the different results be interpreted?

[*] **Correspondence:**
Gabriel Chun-Yeung Lee, University of Nottingham, United Kingdom. ✉ gabriel.lee@nottingham.ac.uk

A critical question arises when ordinal data are used for statistical modelling: *Could ordinal data be treated as continuous data?* Statistical techniques, such as structural equation modelling (SEM), including confirmatory factor analysis (CFA), are often applied to educational research. SEM was developed assuming that the data are continuous (Browne, 1984; Jöreskog, 1967). The technicalities of SEM with continuous data are detailed in many textbooks (e.g., Brown, 2006; Kline, 2016; Schumacker & Lomax, 2004). Theoretically, such a technique is not designed for ordinal data, since computation of means and covariates of ordinal data is not well-defined, unlike continuous data. However, many researchers treat ordinal data as continuous (or interval) in analysis (Robitzsch, 2020; Wu & Leung, 2017); based on normal theory, Pearson correlations and maximum likelihood (ML) estimator are used for computing model estimates. Garson (2015) describes this approach as "do nothing" (p. 253). This article follows the terminologies used in quantitative research literature (e.g., Foldnes & Grønneberg, 2022; Rhemtulla et al., 2012) and refers to the approach that treats ordinal data as ***cont***inuous and uses ***ML*** estimation as *cont-ML*.

Another approach is to add an intermediate process to the standard SEM (Jöreskog, 1994; Muthén, 1984). The process involves estimation of *polychoric correlations* that assume that the ***ord****ered-categorical* data are observed from discretisation of a continuous distribution underlying each variable. Then, model parameters are computed using the polychoric correlations and least squares (***LS***) methods. This article refers to this approach as *cat-LS*. Note the possibility of mixing cat-LS and cont-ML, such as analysing polychoric correlations with ML estimator, and that there exist other estimation methods, such as the Bayesian method (Merkle & Rosseel, 2018) and partial least squares (Hair et al., 2021). However, compared to cat-LS and cont-ML, those strategies have not yet been widely applied (see results); for simplicity, those strategies are not the focus of this article.

This article aims to raise attention to questions and discussions concerning whether ordinal data could be treated as continuous in SEM. It contributes to research by proposing a standard that guides researchers to more transparent reporting of their SEMs with ordinal data. At present, tutorials on SEM/CFA are available in textbooks (e.g., Brown, 2006; Kline, 2016; Schumacker & Lomax, 2004) and online resources (e.g., Putnick & Bornstein, 2016; Rosseel, 2012; Schreiber et al., 2006). However, to the best of my knowledge, those materials primarily focus on SEMs that involve continuous data, rather than ordinal data. Discussion on how ordinal data could be treated in SEM and how to report such SEM appears to be limited. Moreover, there are discrepancies between reporting practices of SEM and the reporting suggested by the existing materials (Schreiber et al., 2006), hindering transparency and reproducibility of SEM reported. Practices and reporting of SEM should encourage transparent and reproducible analyses (LeBeau et al., 2021). Therefore, a standard is proposed in this article to support SEMs, particularly, that involve ordinal data. Rather than turning down either cont-ML or cat-LS, the proposed standard aims to encourage researchers to document decisions made in their SEMs, facilitating reviews by fellow researchers.

This article is divided into four parts. First, an overview of the debate on cont-ML and cat-LS SEM is presented. Then studies that involved SEM and were reported in a mathematics education journal were reviewed. To guide this review, a question was asked: *How were the conduct and results of SEM reported in scholarly publications of mathematics education research?* The results were also compared to studies that involved SEM reported in another educational research journal for reference. This review is novel compared to other reviews on SEM (e.g., Jackson et al., 2009; Schreiber et al., 2006) as it takes the debate on cont-ML and cat-LS into account and reveals practices of reporting of SEM in mathematics education research. Informed by the said debate and review, a standard for reporting SEM with ordinal data is then proposed. This standard aims to support researchers' reflection on transparent and reproducible practices of SEM; reviewers are also encouraged to use this standard for the evaluation of SEM reports. The proposed standard is illustrated by an example of CFA. This article ends with remarks on the review, the standard, and the CFA example.

The Debate

Could ordinal data be treated as continuous in SEM? On the one hand, some researchers prefer cont-ML because: Ordinal data may behave similarly to continuous data under certain conditions (Alabi & Jelili, 2023; Wu & Leung, 2017). Several simulation studies showed that when the number of categories is high enough (e.g., ≥5 categories of an item) and the data are symmetrically distributed across categories, cont-ML can yield reasonably well results (Li, 2016a, 2016b; Rhemtulla et al., 2012). For example, under certain conditions, cont-ML may achieve more acceptable Type I error control and may produce lower standard error estimates, compared to cat-LS (Li, 2016a, 2016b; Rhemtulla et al., 2012). Moreover, cont-ML is less expensive than cat-LS in terms of skills and costs (Robitzsch, 2020). cont-ML is more familiar to researchers, and many textbooks and workshops on cont-ML are available for new users. Computing programmes that support further investigations, such as missing data analysis and testing for random slopes in multilevel model, are more readily developed and available for *all* researchers in cont-ML than in cat-LS (Narayanan, 2012; Robitzsch, 2020).

On the other hand, there are arguments against cont-ML but in favour of cat-LS. First, ordinal data are not continuous, and computation of mean and variance of ordinal data is not well defined. Neither are ordinal data, such as Likert responses, interval in nature, as for example, the interval between 'strongly agree' and 'agree' is not necessarily equivalent to the interval between 'agree' and 'neutral'. This contradicts the assumption of multivariate normality of data in cont-ML, and therefore, cont-ML is theoretically not ideal for ordinal data analysis. Second, whilst cat-LS was not fully accessible in the past, recent advancements have lowered its costs. For example, R package *lavaan* (Rosseel, 2012) has been made available for cat-LS. Recent textbooks do include a section, albeit often shorter, on cat-LS (e.g., Garson, 2015; Kline, 2016). Third, cont-ML often underestimates correlation between two ordinal variables, thereby inaccurate resultant estimation of model parameters (Foldnes & Grønneberg, 2022; Holgado-Tello et al., 2010). On the contrary, cat-LS outperforms cont-ML in most situations. For example, cat-LS can acceptably recover model parameters under most conditions in simulation studies, even when underlying normalities in ordinal data are violated (Foldnes & Grønneberg, 2022; Grønneberg & Foldnes, 2024; Holgado-Tello et al., 2010; Rhemtulla et al., 2012).

Whilst many studies have found cat-LS superior to cont-ML (e.g., Brandenburg, 2024; Holgado-Tello et al., 2010; Rhemtulla et al., 2012), some have shown strengths and weaknesses of cat-LS and cont-ML. For example, despite their outperformance to cont-ML, cat-LS may overestimate correlations, especially when sample size is small and when underlying distributions of the data are non-normal (Foldnes & Grønneberg, 2019, 2022; Li, 2016a, 2016b). Since it is unlikely to detect and make accurate assumptions of underlying distributions (Grønneberg & Foldnes, 2024; Robitzsch, 2020), some researchers have introduced methods of adjusting polychoric correlations by specifying underlying item distributions (Grønneberg & Foldnes, 2024; Kolbe et al., 2021). Doing so allows sensitivity analysis for identifying appropriate model specification.

Acknowledging the debate on cat-LS and cont-ML allows researchers to think critically about strengths and weaknesses of both strategies, both theoretically and practically. Informed decisions could then be made on analysis strategies. As pointed out, cat-LS and cont-ML, as well as other SEM approaches, may yield factor models that disagree; without expert knowledge, appropriate model specification may be hardly identified. Considering academic publication both as a means to disseminate research work and as an invitation to external reviews that helps in improving the work, I argue that reporting of SEM should be made as transparent as possible and should enable reproducible analysis. However, a thorough SEM report would likely increase costs and discourage readers, particularly who have less experience in SEM, from engaging in the report. Therefore, this article aims to support healthy reporting and reviewing of SEM by proposing a standard.

Theoretical Framework

Review of SEM Reporting

To achieve such an aim, practices of reporting SEM in academic publications should be looked at, such that potential weak points of SEM reporting that hinder transparent and reproducible analyses can be reflected, and recommendations can be made. However, only few reviews have focused on reporting of SEM. For example, Schreiber et al. (2006) used 16 articles in *The Journal of Educational Research* to produce a table that compares the information about CFA reported in those articles. Jackson et al. (2009) reviewed 194 articles in psychology journals and compared their reporting practices in CFA to several reporting guidelines, identifying areas of good practices and improvement. Recommendations produced by these two reviews and other associated reviews and guidelines (e.g., Brown, 2006; Kline, 2016; Schumacker & Lomax, 2004) have covered a wide range of elements that encourage transparent and reproducible SEM. However, how ordinal data are treated has not yet been highlighted, despite the frequent use of ordinal variables in educational research and the debate discussed previously. The review reported below therefore draws particular attention to the treatment of ordinal data in SEM whilst being aware of other elements that can encourage transparent and reproducible SEM.

Data Source and Sample

This review explored patterns of reporting of SEM analysis in mathematics education research, guided by the question: *How were the conduct and results of SEM reported in the publications?* Mathematics education provides a good sample space for exploring the use of SEM in an educational research field, because of its wide range of use of SEM (see results). For example, in this field, SEM is often used to validate and explore the factor structure of Likert items or tests (e.g., Lenz et al., 2024), to assess measurement (in)variances (e.g., Zhang et al., 2023) and to analyse longitudinal data (e.g., Street et al., 2022). Since ordinal data are often involved, different strategies are employed to treat data in this field (see results). Moreover, as quantitative literacy is a field of research in mathematics education, and many researchers in this field have a background in mathematics, one would expect that mathematics education researchers have reasonable knowledge of using, reporting, and reviewing quantitative analysis (whilst not being statisticians).

This review used articles published in five mathematics education journals. The selected journals were *Educational Studies in Mathematics*, *The Journal of Mathematical Behavior*, *Mathematics Education Research Journal*, *Research in Mathematics Education*, and ZDM Mathematics Education, because they publish reports of empirical research and are influential in the field. Apart from their similarities, the journals also differ in their affiliation to regional mathematics education associations and composition of editorial boards, suggesting diversity of their audiences. Despite not exhausting all research journals of mathematics education, this review illustrated how SEM has generally been conducted and reported by mathematics education researchers.

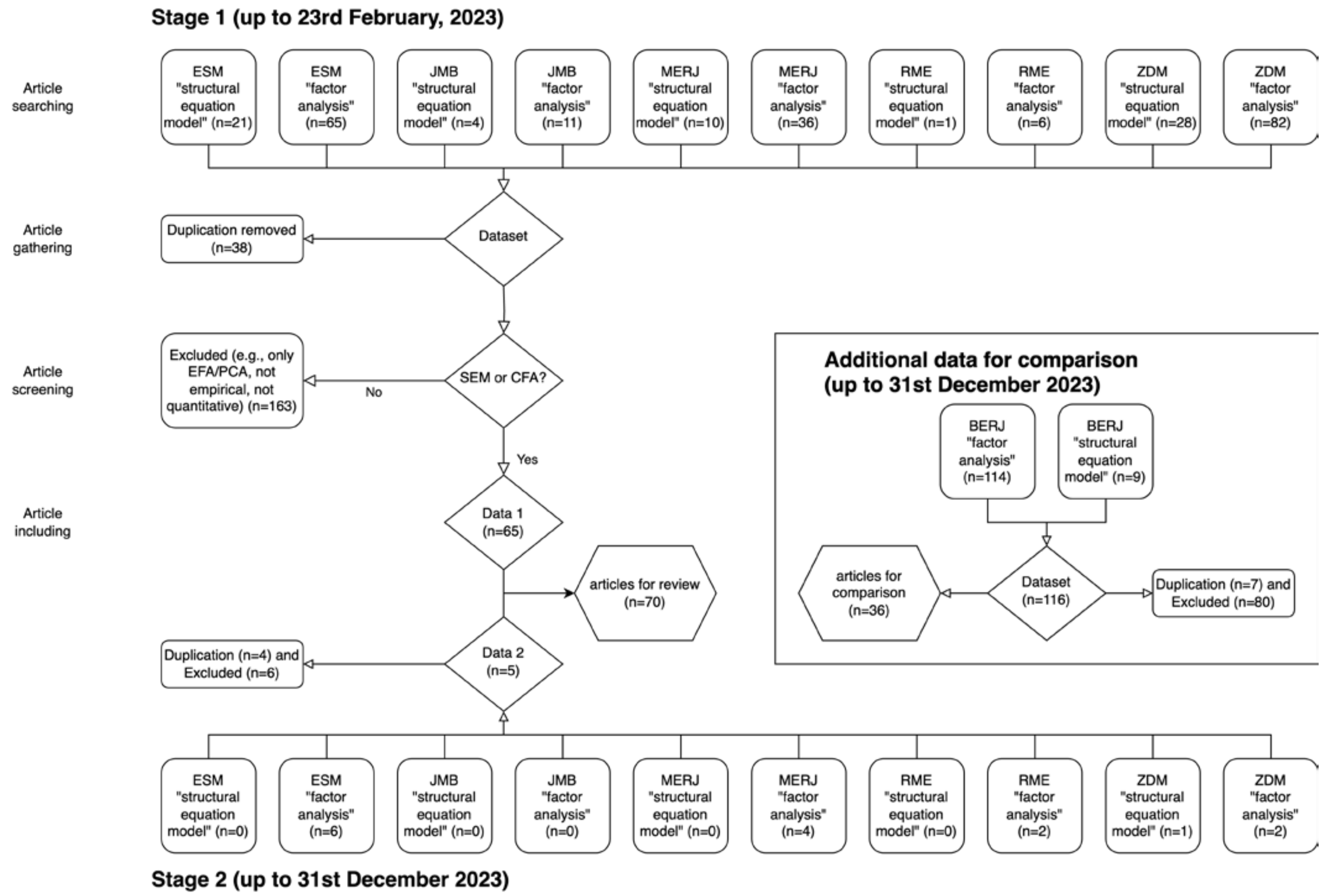
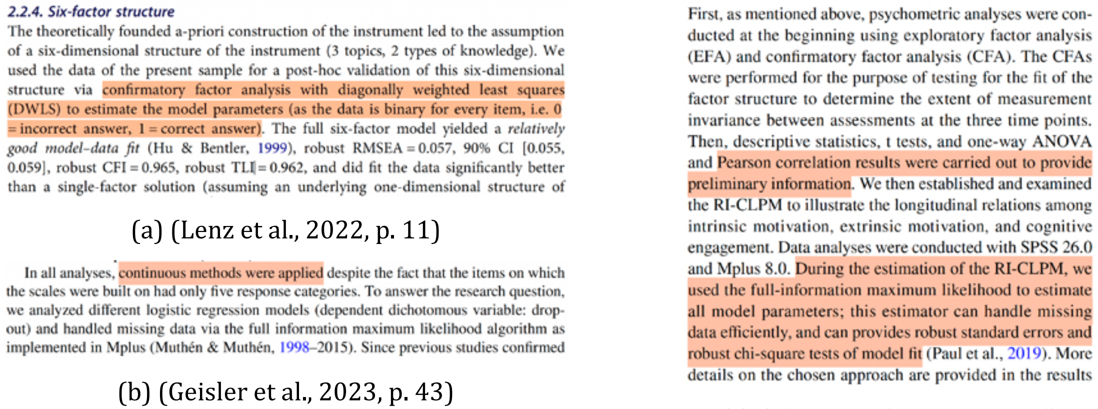

Stage 1 (up to 23rd February, 2023)
Article searching
ESM "structural equation model" (n=21)
ESM "factor analysis" (n=65)
JMB "structural equation model" (n=4)
JMB "factor analysis" (n=11)
MERJ "structural equation model" (n=10)
MERJ "factor analysis" (n=36)
RME "structural equation model" (n=1)
RME "factor analysis" (n=6)
ZDM "structural equation model" (n=28)
ZDM "factor analysis" (n=82)
Article gathering
Duplication removed (n=38)
Dataset
Article screening
Excluded (e.g., only EFA/PCA, not empirical, not quantitative) (n=163)
No
SEM or CFA?
Yes
Article including
Data 1 (n=65)
articles for review (n=70)
Duplication (n=4) and Excluded (n=6)
Data 2 (n=5)
Additional data for comparison (up to 31st December 2023)
BERJ "factor analysis" (n=114)
BERJ "structural equation model" (n=9)
articles for comparison (n=36)
Dataset (n=116)
Duplication (n=7) and Excluded (n=80)
ESM "structural equation model" (n=0)
ESM "factor analysis" (n=6)
JMB "structural equation model" (n=0)
JMB "factor analysis" (n=0)
MERJ "structural equation model" (n=0)
MERJ "factor analysis" (n=4)
RME "structural equation model" (n=0)
RME "factor analysis" (n=2)
ZDM "structural equation model" (n=1)
ZDM "factor analysis" (n=2)
Stage 2 (up to 31st December 2023)

Figure 1. Flowchart of Article Selection Exclusion

I analysed 70 selected articles (see supplementary materials), supplemented with appendices if applicable. The articles were selected from a dataset that was generated by using keywords “structural equation model” and “factor analysis” in each journal. I then scanned the documents in the dataset for indication of the use of SEM/CFA in the study reported, whereas articles that did not involve any empirical studies, or the studies did not involve any SEM/CFA (e.g., involving only exploratory factor analysis, citing a reference that includes “structural equation model”), were excluded (Figure 1).

Review Procedures

The selected articles were analysed according to what (e.g., data nature, estimation methods, fit statistics) and how detailed (e.g., discussions on data distributions, decisions made around analysis) CFA/SEM was reported. For example, when an article contained some statement about the use of ML estimator in the analysis, it was coded as ‘with estimation method (Yes)’; otherwise, ‘No’. To identify how data were treated in an analysis, I searched for statements where author(s) explicitly acknowledged the nature of their data and/or made relevant assumption(s) and decisions. Together with ‘estimation method’, a study was coded as ‘cat-LS’ when author(s) acknowledged the ordinal nature of their data and used LS estimator, a study was coded as ‘cont-ML’ when data were stated to be treated as continuous, and ML estimator was used (for example, see Figure 2). A study could receive multiple codes depending on the analysis methods used. Whilst coded separately, a few studies were grouped into ‘other strategy’ given the low number. A study was coded as ‘unknown strategy’ if author(s) did not explicitly discuss the nature of their data, nor was ‘estimation method’ in SEM/CFA discussed.



2.2.4. Six-factor structure

The theoretically founded a-priori construction of the instrument led to the assumption of a six-dimensional structure of the instrument (3 topics, 2 types of knowledge). We used the data of the present sample for a post-hoc validation of this six-dimensional structure via confirmatory factor analysis with diagonally weighted least squares (DWLS) to estimate the model parameters (as the data is binary for every item, i.e. 0 = incorrect answer, 1 = correct answer). The full six-factor model yielded a *relatively good model–data fit* (Hu & Bentler, 1999), robust RMSEA = 0.057, 90% CI [0.055, 0.059], robust CFI = 0.965, robust TLI = 0.962, and did fit the data significantly better than a single-factor solution (assuming an underlying one-dimensional structure of

(a) (Lenz et al., 2022, p. 11)

In all analyses, continuous methods were applied despite the fact that the items on which the scales were built on had only five response categories. To answer the research question, we analyzed different logistic regression models (dependent dichotomous variable: drop-out) and handled missing data via the full information maximum likelihood algorithm as implemented in Mplus (Muthén & Muthén, 1998–2015). Since previous studies confirmed

(b) (Geisler et al., 2023, p. 43)

First, as mentioned above, psychometric analyses were conducted at the beginning using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The CFAs were performed for the purpose of testing for the fit of the factor structure to determine the extent of measurement invariance between assessments at the three time points. Then, descriptive statistics, t tests, and one-way ANOVA and Pearson correlation results were carried out to provide preliminary information. We then established and examined the RI-CLPM to illustrate the longitudinal relations among intrinsic motivation, extrinsic motivation, and cognitive engagement. Data analyses were conducted with SPSS 26.0 and Mplus 8.0. During the estimation of the RI-CLPM, we used the full-information maximum likelihood to estimate all model parameters; this estimator can handle missing data efficiently, and can provides robust standard errors and robust chi-square tests of model fit (Paul et al., 2019). More details on the chosen approach are provided in the results

(c) (Zhang et al., 2023, p. 405)

Figure 2. Examples of Studies Coded as cat-LS (a) or cont-ML (b & c)

Given that it involved primarily identification of discussions on aspects of SEM in the articles, the analysis did not involve any second coder, but two strategies have been applied to confirm the results: the articles were reviewed and coded multiple times at different time points with absolute agreement of codes[i] ranging between 84% and 99%, and comments on the codes were recorded and reviewed to resolve and finalise disagreeing codes. The results were then compared with 36 studies in *British Educational Research Journal* (BERJ), as *ad hoc* information, to link the results to general educational research. For the ease of communication, the 'review' (or 'reviewed' studies) refers to the mathematics education articles, as the BERJ articles are considered supplementary.

Results

Researchers had divergent preferences for what information about SEM to report. On the one hand, most articles discussed factor or path models tested (87%), model selection criteria (79%), model fit statistics (94%) and software used for analysis (84%). On the other hand, fewer articles discussed descriptive statistics of data (43%), how data were treated in analysis (49%), item correlation matrix (31%) and model estimation algorithm (43%). Only a few articles provided rationale behind the choice of model estimation algorithm (27%) and technical information about model specification (14%).

Table 1. Information About SEM Being Reported

Code Description	**Percentage of Reviewed Articles (n=70)**	**Percentage of BERA Articles (n=36)**
Model fit statistics (e.g., chi-square) being reported	94%	94%
Factor or path model(s) tested being stated	87%	75%
Software used for analysis being stated	84%	67%
Model selection criteria (e.g., non-significant chi-square) being discussed	79%	64%
Model parameter estimates being reported	76%	89%
Nature of observed variables (e.g., as ordinal) being acknowledged and/or assumptions (e.g., to be continuous) being discussed	49%	36%
Descriptive statistics of observed variables (e.g., frequency table, skewness, kurtosis) being reported	43%	33%
Model estimation algorithm being stated	43%	39%
Missing values and how they were treated being stated	31%	44%
Table(s) of correlation matrix of observed variables being presented	31%	28%
Rationale behind choice of model estimation algorithm (e.g., "XXX was used because/so that YYY") being stated	27%	17%

Table 1. Continued

Code Description	Percentage of Reviewed Articles (n=70)	Percentage of BERA Articles (n=36)
Technical information about model specification (e.g., programme scripts, model re-specification process) being discussed	14%	14%
Appendices or supplementary materials are being used to provide extra information about SEM/CFA	14%	25%

Note. Code descriptions are arranged in descending order of the percentage of reviewed articles coded

All reviewed studies involved ordinal data with different strategies to analyse the data: 26% used cont-ML, compared to only few cat-LS (7%) and other strategies (7%). Yet in most of the articles (60%), strategies applied to the study could not be identified. Of which, 7 hinted at cont-ML by reporting use of ML estimator in analysis, it remained unclear whether cont-ML was used as polychoric correlations could be analysed with ML estimator. Similarly, another 7 articles, that reported conversion of ordinal item scores to variables with more scale points or Pearson correlations between variables, were not identified as cont-ML because Pearson correlations could be analysed with LS estimator, which can yield unreliable and overoptimistic results (Olsson et al., 2000).

Table 2. Types of Treatment of Ordinal Variables

Code Description	Percentage of Reviewed Articles (n=70)	Percentage of BERA Articles (n=36)
Unknown: not enough information to identify how ordinal data were treated in analysis (e.g., missing of information about estimation method, acknowledgement or assumption of the nature of data, or neither)	60%	72%
cont-ML: maximum likelihood method being used to estimate model parameters, data acknowledged/ assumed and treated as continuous (e.g., use of Pearson correlations)	26%	17%
cat-LS: least squares method being used to estimate model parameters, data acknowledged and treated as ordinal (e.g., use of polychoric correlations)	7%	8%
Other strategy: e.g., Bayesian method, partial least squares method	7%	3%

Note. Code descriptions are arranged in descending order of the percentage of reviewed articles coded

Patterns of the reporting of BERJ articles were consistent with those in the mathematics education research journals (Table 1), except that the software used for analysis was arguably less often reported in BERJ articles (67%) than in the reviewed articles (84%). The Fisher's exact test was significant p = 0.048, and the Chi-square test of independence was marginally significant $\chi^2(1) = 3.369$, p = 0.066. This difference also reflected a slightly lower amount of BERJ articles being identified as cont-ML (17%) than in the reviewed articles (26%), although the difference was not statistically significant; $\chi^2(1) = 0.655$, p = 0.418. Moreover, 4 out of the BERJ articles which strategy was unknown (72%) hinted at cont-ML by reporting model parameters being estimated with ML estimator. The consistency in how SEM studies were reported at mathematics education research journals and BERJ gave a more confident result of the reporting of SEM in educational research.

Conclusion

The review showed different choices of SEM information that researchers decided to report. Many provided model parameters and fit statistics to justify the final model(s) and results of their analysis, but only few provided enough information to guide readers through the estimation strategy of model parameters, especially the treatment of ordinal variables. LeBeau et al. (2021) argued that research can benefit from

> "(1) increased transparency and trust that the study results are reported accurately, (2) increased ability to know the analytic steps exactly though analytic code by releasing the analysis code, and (3) increased ability to explore the raw data through open access to the data used on the analysis" (p. 197).

Data (underlying) distribution can impact SEM too. Less than half of the reviewed articles presented descriptive statistics, including frequencies, kurtosis, skewness, range, mean and standard deviation, and/or acknowledged non-normality of data, for example, failing of normality tests. Given that simulation studies have shown that response (underlying) distribution can impact parameter estimation of SEM (Foldnes & Grønneberg, 2022; Li, 2016a; Rhemtulla et al., 2012), I argue that presenting data distribution is essential as it enables readers' evaluation of researchers' decisions on ordinal data analysis (e.g., cat-LS or cont-ML, applying adjustments or not).

Given the issues raised by inconsistent practices of reporting of SEM, some scholars have suggested what to report, such as software programme(s) used, correlation matrix, sample size, means and standard deviations of variables, and theoretical model(s) tested (e.g., Kline, 2016; Schumacker & Lomax, 2004). However, those suggestions do not include issues raised by inclusion of ordinal data. I argue that it is vital for researchers to discuss the decisions that they make when conducting SEM with ordinal variables. Doing so enables readers to replicate the analysis if data (or correlation/covariance matrix) are provided (LeBeau et al., 2021; Stodden, 2015). Readers may also use the information to re-construct and compare cat-LS and cont-ML results, with reasonable caution.

Next, a standard framework is proposed to guide conduct and reporting of SEM with ordinal variables, illustrated by an example of CFA. The standard contributes to research by adding the type(s) of variables and data treatment into consideration in SEM reporting, and the example adds to research by using real-life research data (rather than simulated data) to show that different treatments can lead to results that disagree. In supplementary materials, I also demonstrate that the standard can guide reconstruction of CFA, through reported correlation/covariance matrix and data distribution (or data matrix), estimation algorithm, and model configuration. Such reconstruction of an analysis can encourage validating the analysis being reported. This is compared to the findings from the review, that most of the reviewed articles did not include some of the said information, thereby reducing reproducibility of the analysis. Together, this article argues for the importance of transparent and reproducible SEM reporting.

The Proposed Standard

This standard extends existing scholarly suggestions on SEM (e.g., Kline, 2016; Schumacker & Lomax, 2004) and statistical analysis in general (e.g., LeBeau et al., 2021; Stodden, 2015; Wright, 2003). The key principle is to provide necessary information that can guide readers of SEM report through decisions made around and reconstruction of analysis. I present this standard in the form of reflection points (or a checklist, see Appendix), accompanied by an example of CFA as an illustration. Being part of the data from a study on pre-service teachers' beliefs about proof in school mathematics (Lee, 2022), the dataset was used to evaluate 22 Likert items that aimed to measure pre-service teachers' beliefs about proof. Table 3 shows the statements and response distributions of the items.

Table 3. Likert Items Measuring Pre-Service Teachers' Beliefs about Proof (Adapted from Keçeli-Bozdağ et al., 2015)

Item	Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Missing
x1	I feel myself under pressure when I make proofs in mathematics lessons.	8	84	97	48	3	1
x2	I do not like dealing with mathematical proofs.	9	39	68	102	22	1
x3	It's fun for me to make proofs.	29	122	67	15	7	1
x4	One of the things I like about mathematics is the proofs of theorems in the lessons.	19	101	75	36	8	2
x5	I do not worry when making proofs.	9	95	80	47	9	1
x6	Making proofs arouses my curiosity.	31	128	52	20	8	2
x7	I find it boring to make proofs.	11	36	63	110	19	2
x8	I think that theorems and proofs are the foundations of mathematics.	73	136	27	3	1	1
x9	My self-confidence diminishes when I cannot prove a theorem.	18	112	70	35	4	2
x10	I find it interesting to make proofs.	26	126	67	14	6	2

Table 3. Continued

Item	Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Missing
x11	Proofs are very important in understanding mathematical language.	59	148	28	5	0	1
x12	Proofs are indispensable for mathematics.	85	129	24	2	1	0
x13	Making proofs improves mathematical thinking.	80	138	15	7	1	0
x14	To me, it is important to make proofs in mathematics.	56	147	32	6	0	0
x15	Working on a proof in front of the class frightens me.	12	50	104	67	7	1
x16	I enjoy proving mathematical results.	30	138	59	9	5	0
x17	To improve logical thinking, making proofs is necessary.	33	147	53	7	1	0
x18	I feel enthusiastic to prove a theorem when I see it.	12	66	108	44	10	1
x19	I love mathematics, but I do not like making proofs.	19	43	93	77	9	0
x20	In the development of deductive reasoning, making proofs plays an important role.	38	143	55	5	0	0
x21	I cannot see the point of making proofs that have already been proven beyond doubt by mathematicians and scientists.	12	34	61	101	33	0
x22	Not being able to prove upsets me.	23	113	67	33	5	0

n=241

What Model(s) are Involved?

The first question concerns models to be tested. In SEM, as a method for studying relationships between variables, a sound theoretical foundation should be established (Schumacker & Lomax, 2004). This includes a discussion on the rationale and purpose of the study. Models to be tested should be discussed concerning their theories, which can also be coupled with discussion on data collection instruments. For example, during selection of instruments, questions can be asked, such as '*how many scale points of an item would suffice to measure both commonality and diversity of participants?*' and '*how suitable are variables with different units and point scales for the model to be tested?*' Decisions on instruments, such as item selection, and scaling, can be accounted for with literature, purpose of study and models to be tested. Not only can such discussion justify the models and instruments, but it can also make a case for treatment of data in subsequent SEM.

In the example, CFA was used to test the factor structure of the items. Backgrounds of the items formed the foundation of this CFA and the models tested. The items were adapted from Keçeli-Bozdağ et al. (2015), which were in turn based on research (e.g., Almeida, 2000; Kotelawala, 2007; Nyaumwe & Buzuzi, 2007). For the sake of using the items with pre-service teachers in Hong Kong, the items, originally written in Turkish, were translated into Chinese and English (for the translation process, see Lee, 2022). It was considered that the translated items were semantically consistent with the original items. The original items measured four dimensions of beliefs about proof (Keçeli-Bozdağ et al., 2015): views/perspectives towards proof, the feeling/emotion one has while proving, common negative attitudes, and negative attitudes during proving (translated by Zengin, 2017). Whilst the original items asked participants to choose from three categories: 'agree', 'undecided/don't know', and 'disagree', the translated items followed 5-point-scaling, from 'strongly agree' to 'strongly disagree', as this setting allowed more precise measurement of how much a participant agreed with a statement. The use of 'neutral' option in this setting avoided forcing the pre-service teachers into expressing agreement or disagreement.

Despite the evaluation of the original items by Keçeli-Bozdağ et al. (2015), CFA was still needed because, first, it was unclear whether the original and the translated items had the same factor structure. It is possible that, due to cultural and educational differences, Turkish and Hong Kong pre-service teachers interpreted and responded to the items differently, resulting in different factor structures observed from the 'same' items (Andrews & Diego-Mantecón, 2015). Second, given that the number of scale points can impact SEM, it was unclear how the use of 5-point-scaling, rather than 3-point-scaling, yielded a CFA solution inconsistent with the principal component analysis by Keçeli-Bozdağ et al. (2015).

Three models were initially tested and compared in this CFA. Figures 3–5 illustrate the three models, respectively; in each figure, a rectangle represents an observed variable (i.e., an item), an ellipse represents a latent variable, a one-headed arrow represents a regression path between two variables (in the case between an observed variable (head) and a latent variable (end), the path is a factor loading), and a two-headed arrow represents a covariate/correlation path between two variables. The first model (Model 1) included four latent variables, according to Keçeli-Bozdağ et al. (2015).

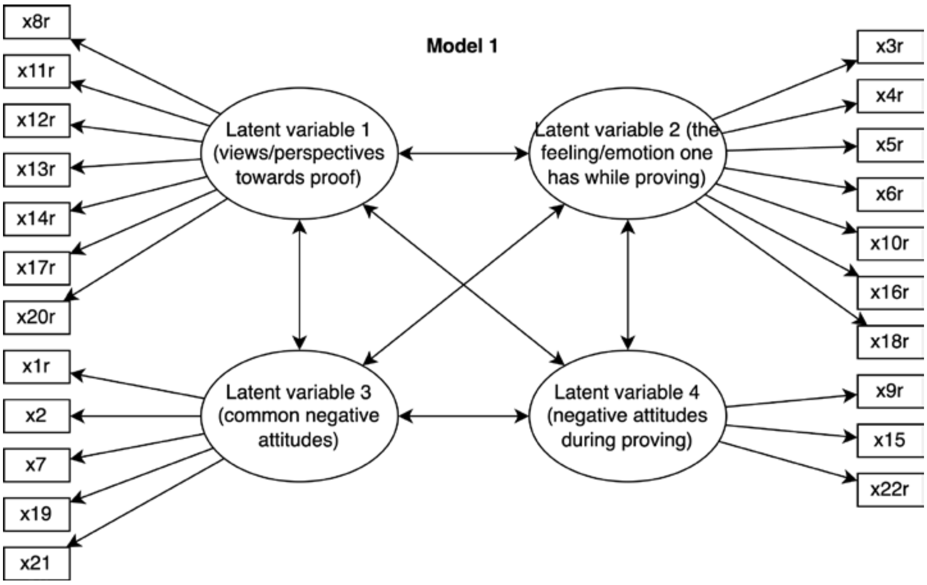

Model 1
x8r
x11r
x12r
x13r
x14r
x17r
x20r
Latent variable 1 (views/perspectives towards proof)
Latent variable 2 (the feeling/emotion one has while proving)
x3r
x4r
x5r
x6r
x10r
x16r
x18r
x1r
x2
x7
x19
x21
Latent variable 3 (common negative attitudes)
Latent variable 4 (negative attitudes during proving)
x9r
x15
x22r

Figure 3. 4-Factor of Pre-Service Teachers' Beliefs about Proof

Model 2 assumed that all items loaded onto a single latent variable, testing whether all items measured such a unidimensional variable, namely, beliefs about proof. Model 3 built on research, in which beliefs about the importance of proof, enjoyment and interest in proof, and anxiety about proof have been found to be major components of beliefs about proof that influence teachers' practice of teaching proof (e.g., Frasier, 2010; Knuth, 2002; Kotelawala, 2007). The major difference between Model 1 and Model 3 was the (re)arrangement of the items and latent variables relating to emotions, such as enjoyment and anxiety.

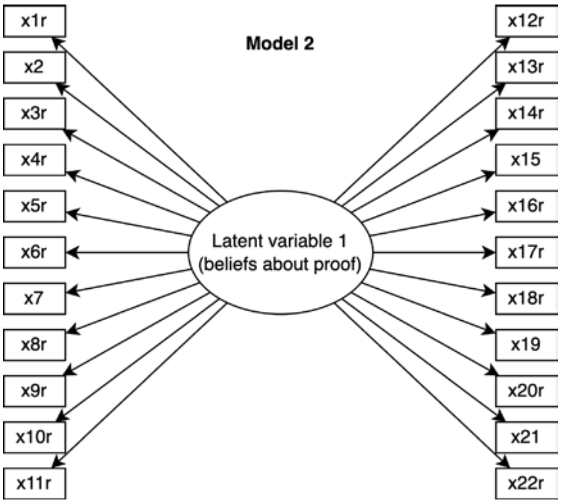


Figure 4. 1-Factor of Pre-Service Teachers' Beliefs about Proof

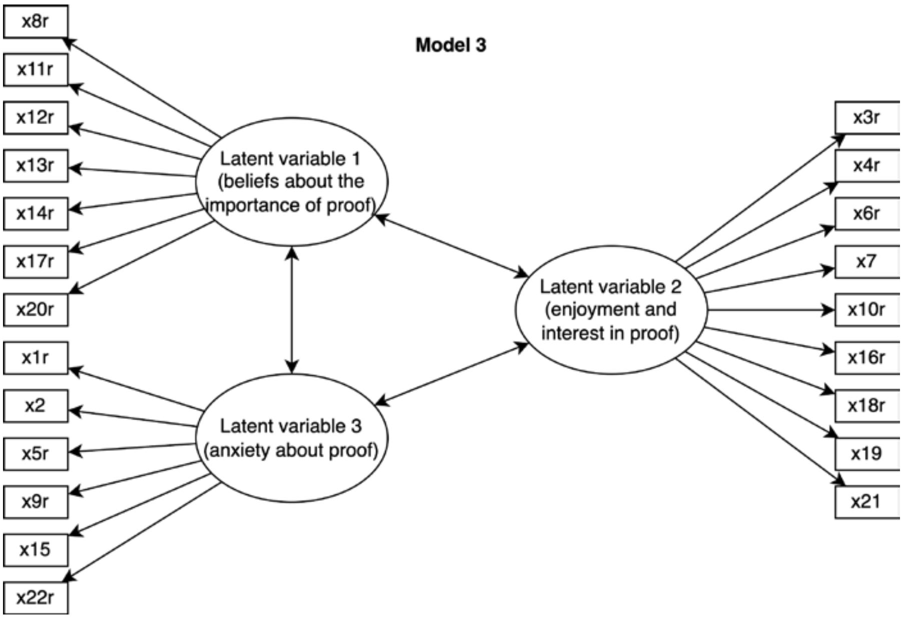

Model 3
x8r
x11r
x12r
x13r
x14r
x17r
x20r
Latent variable 1 (beliefs about the importance of proof)
x1r
x2
x5r
x9r
x15
x22r
Latent variable 3 (anxiety about proof)
Latent variable 2 (enjoyment and interest in proof)
x3r
x4r
x6r
x7
x10r
x16r
x18r
x19
x21

Figure 5. 3-Factor of Pre-Service Teachers' Beliefs about Proof

What type(s) of data are involved?

The second question concerns data being collected. Researchers should discuss observed and latent variables, and how they are defined in analysis. Particularly, researchers should account for how each ordinal variable is treated, what measures are implemented for any missing values, and how any violations of assumptions for SEM are evaluated.

The following questions aim to guide reflection on the nature and treatment of data. *How are observed variables scored? Is each observed variable a score given to a participant's response to one single item, for example, a 4 when a participant chooses 'strongly agree' for a statement? If a variable is a score given to responses to several items, for example a test of numeracy, how is the score computed? Is it computed by summing 'correct' items, taking an arithmetic mean, using item response theory to generate a plausible score, or other methods?* Relatedly, *how are the scores treated? Are they considered as continuous, for example, given that the number of scale points is large enough? Or are the scores treated as ordinal – and if so, is underlying normality assumed?* If polychoric correlations are used and each of the observed variables is assumed to have non-normal underlying distribution, *is such assumption supported by any existing research? Are those assumptions, decisions and rationale documented?* Answers to these questions often inform decisions on choice of analysis methods.

Furthermore, in practice, having a complete dataset for analysis is rare due to withdrawals from study and incomplete responses to some items. Missing values can impact SEM results, sometimes drastically, when missingness is high (Kline, 2016). Therefore, a good practice is to record any missingness and discuss relevant measures (e.g., multiple imputation of missing values) justified by literature (for more information, Cañete-Massé et al., 2022; Kline, 2016; Schumacker & Lomax, 2004). Issues around treatment of ordinal variables and missingness are highly relevant to the next reflection point, violation of SEM assumptions.

Assumptions for SEM depend on estimation methods used in computation. For example, assumptions for bivariate correlation estimation apply to SEM. If Pearson correlations are used, then it is assumed that the data are continuous and multivariate-normally distributed (Kline, 2016). If polychoric correlations are used, then ordinal data are assumed to be discretised from some underlying (normal) distributions (Muthén, 1984). Similar ideas apply to model estimation algorithms. However, in practice, data do not always meet all assumptions for SEM. For example, when a test developed for 8-year-olds is implemented with 15-year-olds, the distribution of test scores is likely skewed towards full mark, not normal. Violations of assumptions can impact SEM results to varying degrees (Khine, 2013; Kline, 2016). Researchers can record any assumption violations observed and discuss decisions on measures implemented (e.g., robust corrections of fit statistics). Different measures and sensitivity analysis can also be considered to test for robustness of results (Grønneberg & Foldnes, 2024; Kline, 2016).

By reflecting on the nature of data, researchers can evaluate their choice of SEM methods. For example, if data are highly non-normal, it is less likely that cont-ML would yield reliable results without any measures to address violation of normality (Foldnes & Grønneberg, 2022; Li, 2016a, 2016b; Rhemtulla et al., 2012). If data are binary, tetrachoric correlations might be considered, rather than Pearson correlations, as the former are designed for estimating correlations between binary variables.

In the example, the Likert items collected ordered-categorical responses, which were converted into numerals, from '4=strongly agree' to '0=strongly disagree'. As shown in the models (Figures 3–5), each item was considered individually as an observed variable. Since this article discusses the debate on whether ordinal data could be treated as continuous, both cat-LS and cont-ML were considered in the analysis. Given that most observed item distributions were skewed in this example (Table 3), consistent with existing research (e.g., Almeida, 2000; Frasier, 2010; Nyaumwe & Buzuzi, 2007), I argue that the data were not necessarily normally distributed nor assumed to have underlying normality. Therefore, cat-LS was divided into two conditions: 'ordinal data with ***norm***al underlying distributions' (cat-LS-norm) and 'ordinal data with ***non-norm***al underlying distributions' (cat-LS-nonnorm), and a total of 9 conditions (3 models × 3 treatment conditions) were tested. Moreover, since analysis tends to yield more biased results when having item distributions skewed in opposite directions (Grønneberg & Foldnes, 2024), the items that had an observed distribution skewed towards 'agree' were *reversely* scored. As a result, all observed distributions in the analysis were either symmetrical or positively skewed (towards '0').

There were missing values in the dataset. Inspection of the raw data suggested that responses were missing completely at random, because respondents overlooked some items. Given that missingness was low, incomplete responses (n=7) were omitted from analysis, for simplicity.

What analysis methods are involved?

Different methods can yield different results. Presenting decisions on what methods (e.g., cat-LS or cont-ML, with or without adjustments of estimated values) enables readers to review decisions made by researchers and the resultant SEM outputs. Researchers can ensure research quality and preserve research integrity by documenting and reflecting on decisions on analysis methods; they may also re-explore, and potentially, re-discover interpretations of their SEMs (LeBeau et al., 2021; Stodden, 2015), further contributing to research.

The following questions aim to guide reflection on what to record and discuss regarding analysis methods. *How is treatment of data, regarding type(s) and missingness, reported? Are sampling and sample size discussed?* As sampling and sample size impact result interpretation, reporting them enables readers to inspect relevant claims being made. *Are any violations of SEM assumptions reported, and related measures presented? What software is used?* As different software and versions can yield different numerical outputs due to varying uses of rounding and computation methods, although often slightly, it is worth reporting which software is used. Relatedly, as different estimation methods can lead to different outputs, *are the decisions on estimation methods discussed?* Since SEM is a technique of testing whether a hypothesised model fits empirical data, acceptance or rejection of the model becomes one key outcome of SEM. *How is model fit evaluated (e.g., model selection or rejection criteria)?* For example, if goodness-of-fit indices are used, *what values are considered acceptable, and how is such decision justified?* Moreover, it is worth reflecting that most suggested cut-off values of goodness-of-fit indices are intended as rules of thumb for selecting or rejecting models with continuous, normal data (Savalei, 2021; Savalei & Rhemtulla, 2013). *How valid are those values applied to ordinal data, particularly in cat-LS?* Moreover, given that SEM involves numerical methods of parameter estimation, when analysing ordinal data, particularly with 'not large enough' samples, computational issues, such as non-positive definite covariance matrix and non-convergence, may arise (Flora & Curran, 2004; Li, 2016a, 2016b). When such issues arise, the validity of SEM solution should be cautioned.

Quantitative methodologists suggest a SEM report should include statistics, such as correlation/covariance matrix, goodness-of-fit indices, estimated path parameters and factor loadings (e.g., Brown, 2006; Kline, 2016; Schumacker & Lomax, 2004). If any computational issues occur, such as non-reasonable parameter estimations, non-positive definite covariance matrix and non-convergence, notes should be made, and measures implemented for addressing them should be discussed. Since, often, not all goodness-of-fit indices meet model selection or rejection criteria, discussing decisions on which indices meet the criteria and which do not and whether the tested model(s) are rejected enables readers to evaluate the decision process and the resultant interpretations and implications of a study. If model modification is involved, it is worth discussing decisions on the process, such as use of model modification indices and criteria for item retention or removal. The modified model(s) could then be evaluated.

In cat-LS, ordinal data are assumed to be generated from discretisation of underlying, usually normal, distributions. Thresholds of the discretisation are estimated by dividing an assumed distribution into the number of categories of an item proportional to the actual distribution of the categorical responses. Different assumptions of the 'shape' of underlying distributions can yield different threshold estimations, and hence different bivariate correlations and SEM solutions (Grønneberg & Foldnes, 2024). It is therefore worth discussing, for example, what distribution parameters are assumed. Researchers may also consider sensitivity analysis by applying multiple assumptions of underlying distributions (Grønneberg & Foldnes, 2024). In addition, when linking item responses to underlying distributions, there are two ways of scaling the laten response variables. The choice of parameterisation methods should depend on the purpose of SEM. Whilst interpretation of SEM solutions from delta parameterisation can be 'familiar and straightforward' (Kline, 2016, p. 327), theta parameterisation enables multigroup SEM to test for invariances (Millsap & Yun-Tein, 2004).

Is software code available?

This code can clarify analysis process if discussions have not been made clear enough for readers. This supplement also allows readers to inspect and replicate the analysis process. The code could include: (1) which software packages, including their versions, are used, (2) treatment of missing values, (3) how observed and latent variables of tested model(s) are defined and treated, including assumed (underlying) distributions of the variables, (4) whether and how any parameters of the model(s) are constrained, and (5) estimation methods.

In the example, the working dataset had a sample size of n=234, after incomplete responses were excluded listwise in MS Excel. Given the scope of this CFA as an example of conducting and reporting SEM with ordinal variables, analyses of multigroup invariances (e.g., by gender) were not included. CFA with the working dataset as a single group was conducted using R (Version 4.3.1; R Core Team, 2023) with *lavaan* package (Version 0.6-15; Rosseel, 2012) on macOS.

The three said models were tested (Figures 3–5); for each model, three conditions of data treatment were considered. For cont-ML, data were assumed continuous, and Pearson correlations used for estimation of model parameters. ML estimation with robust corrections (MLR) was used; the corrections were performed to address the skewness of the data and adjust resulting numerical outputs (Savalei, 2014). For cat-LS-norm, ULS estimation with robust corrections of standard errors and mean and variance adjusted test statistic (ULSMV) was used (Savalei, 2014; Savalei & Rhemtulla, 2013) with R code **ordered=TRUE**; that is, the data were identified as ordered-categorical, and polychoric correlations were estimated with an assumption that the data were discretised from normal distributions. Whilst this assumption of latent normality addressed mismatch between 'non-continuous' data and application of analysis for continuous data, the resultant polychoric correlations were still restricted to latent normality (Robitzsch, 2020). Note that the corrections applied to cat-LS adjusted not for non-normality (Savalei, 2014, p. 158). For cat-LS-nonnorm, the polychoric correlation and associated asymptotic covariance matrices of the dataset were adjusted following the method described in Grønneberg and Foldnes (2024), before ULSMV was performed. This adjustment assumed that the middle option 'neutral' sits around 'zero' of the latent response distribution and the transitions between 'strongly agree' and 'agree' and between 'strongly disagree' and 'disagree' are symmetric (see supplementary materials). Given that it was implausible to detect the true distributions of ordinal data with the data alone (Grønneberg & Foldnes, 2024), the three said conditions of data treatment enabled examination of consistency and inconsistency in the CFA model outputs. Since the focus of this example was neither variation in the scaling of latent response variables (x1*–x22*) nor testing for multi-group invariances, delta parameterisation (by default in *lavaan* or with R code **parameterization="delta"**) was used.

The models were evaluated using goodness-of-fit indices: adjusted model chi-square (χ^2), standardised root mean square residual (SRMR), root mean square error of approximation (RMSEA) and comparative fit index (CFI). χ^2 tests whether the predicted model (by parameter estimates) differs from the observed data (by correlation matrix); a significant χ^2 indicates the two are different. SRMR evaluates the difference between the predicted model and the observed data by subtracting their correlation matrices and taking the square root of the mean square. Between 0 and 1, a SRMR value close to 0 indicates the difference between the model and the data is minimal. RMSEA takes model parsimony into consideration by using χ^2 and the number of freely estimated parameters (degrees of freedom, df), and 'punishes' a model for poor model parsimony, testing the extent to which the model fits reasonably well in the data. Usually between 0 and 1, a RMSEA value close to 0 indicates the difference between the model and the data is minimal. CFI evaluates the model in relation to a null model, in which all relationship paths between the observed variables are fixed to be zero but the variance of each variable is estimated. Between 0 and 1, a CFI value can be interpreted as how much the fit of the model is better than that of the null model; the closer to 1 the better. There are many other goodness-of-fit indices available in literature and software, but these four are the most common. Incorporating multiple indices enables evaluation of model fit from different aspects, as a form of triangulation (Brown, 2006; Kline, 2016). However, different indices have their own weaknesses; for example, χ^2 is sensitive to sample size and normality of data. Equations and formulae of computation are overviewed in supplementary materials; interested readers are advised to explore the references for technicalities.

Based on common cut-off criteria, non-significant χ^2, RMSEA,SRMR≤0.08 and CFI≥0.95 indicate 'good fit' whereas significant χ^2, RMSEA,SRMR>0.1 and CFI<0.9 indicate 'poor fit' (Brown, 2006; Kline, 2016), for simplicity. It is, however, worth recalling that these criteria are originally developed for cont-ML under ideal conditions; their suitability for cat-LS has not yet been widely studied (Savalei, 2021). Therefore, cat-LS goodness-of-fit statistics should be interpreted with extra care along with additional arguments.

Table 4. Goodness-of-Fit Statistics of the Models

Treatment	Model	χ^2	df	p-value	RMSEA	SRMR	CFI
cont-ML	1	433.782	203	0.000	0.076	0.082	0.879
cont-ML	2	848.636	209	0.000	0.126	0.123	0.656
cont-ML	3	482.760	206	0.000	0.083	0.085	0.853
cont-ML	4	276.147	193	0.000	0.047	0.051	0.956
cont-ML	5	94.734	91	0.374	0.015	0.034	0.997
cat-LS-norm	1	700.528	203	0.000	0.128	0.089	0.776
cat-LS-norm	2	1327.948	209	0.000	0.172	0.140	0.580
cat-LS-norm	3	715.989	206	0.000	0.134	0.092	0.750
cat-LS-norm	4	345.756	193	0.000	0.094	0.054	0.884
cat-LS-norm	5	112.782	91	0.061	0.068	0.035	0.959
cat-LS-nonn	1	726.181	203	0.000	0.118	0.086	0.782
cat-LS-nonn	2	1370.783	209	0.000	0.159	0.134	0.591
cat-LS-nonn	3	748.681	206	0.000	0.125	0.089	0.754
cat-LS-nonn	4	356.245	193	0.000	0.086	0.052	0.891
cat-LS-nonn	5	118.035	91	0.030	0.064	0.034	0.959

n=234

Goodness-of-fit statistics and computational issues indicated that all three models were poor fit under all three data treatment conditions. The goodness-of-fit statistics of Models 1–3 did not reach the said cut-off criteria (Table 4). For example, cont-ML results showed that difference between Model 1 and the data was statistically significant: $\chi^2(203)=433.782$, $p<0.001$. Only 88% of the fit of Model 1 was better than that of the null model (CFI=0.879). However, RMSEA(=0.076) and SRMR(=0.082), with a value close to 0.08, hinted Model 1 might fit the data 'marginally well' *when the ordinal data were considered to be continuous*. On the other hand, *when polychoric correlations and ULSMV were used* (i.e., cat-LS-(non)norm), poorer goodness-of-fit statistics showed that Model 1 did not fit the data well (Table 4). Model 1 also yielded problematic values, such as standardised factor loading ≥1 and negative variance of a variable, when cat-LS-(non)norm was used. Overall, the results suggested Model 1 was likely mis-specified, and the factor structure of the translated items was not as same as what Keçeli-Bozdağ et al. (2015) suggested. Likewise, the results showed that Models 2–3 were poor fit.

Moreover, a warning of non-positive definite variance-covariance matrix of estimated parameters occurred in all models when robust corrections (ULSMV) were implemented. Whilst this could be considered as a symptom of unidentified models, the fact that such a warning disappeared when 'naïve' standard errors were not adjusted (**se="standard"**) suggested that the issue might have been caused by the adjustment of standard errors, rather than unidentified models. R code and *lavaan* outputs are available in supplementary materials.

Together, the results suggested that the models should be re-specified to explore a model that better described the data. To balance the purposes of this article and this example, the modification process is recorded in supplementary materials. Only the result is reported here:

Drawing on theories and inspection of the items and previous results, Models 4–5 were developed from Model 3 under cont-ML (Model 4), cat-LS-(non)norm (Model 5), respectively. cont-ML and cat-LS approaches led to different 'final' models after modification. Figures 6–7 illustrate the two models, respectively, with numeric outputs under cat-LS-nonnorm. The figures are presented in the same manner as Figures 3–5, with additional information: for each observed variable, an additional ellipse links between the variable and its latent variable(s) and represents the assumed underlying distribution of the variable (in the case of cat-LS-nonnorm, a gamma Γ distribution with values of shape and scale). The value attached to the underlying distribution with a one-headed arrow represents the unique variance of the variable (error term). Values are also attached to paths of factor loadings and correlations, with standard errors (s.e.) in brackets. For example, in Model 4, x8 had an underlying gamma distribution with shape=0.65 and scale=0.80 and loaded on Latent Variable 1 with a factor loading of 0.57 and a unique variance of 0.67; Latent Variables 1 and 2 had a positive correlation of 0.54.

Based on all 22 items, Model 4 had four latent variables: beliefs about the importance of proof ('importance'), enjoyment and interest in proof ('enjoyment'), anxiety about proof ('anxiety'), and negative reactions when failing to prove ('failing'). Whilst Model 4 achieved 'good fit' under cont-ML, the goodness-of-fit indices under cat-LS approaches rejected the model (Table 4). This showed that Model 4 was a 'better' model of describing the data than Models 1–3, but the model was still considerably different from the data *when polychoric correlations and ULSMV were used*. In the model, the significant correlation (=0.51; s.e.=0.06) between the error terms of x3 and x4 (Figure 6) suggested an unexplained effect, presumably, of consecutive, similar items; in other words, this may be an indication of item redundancy.

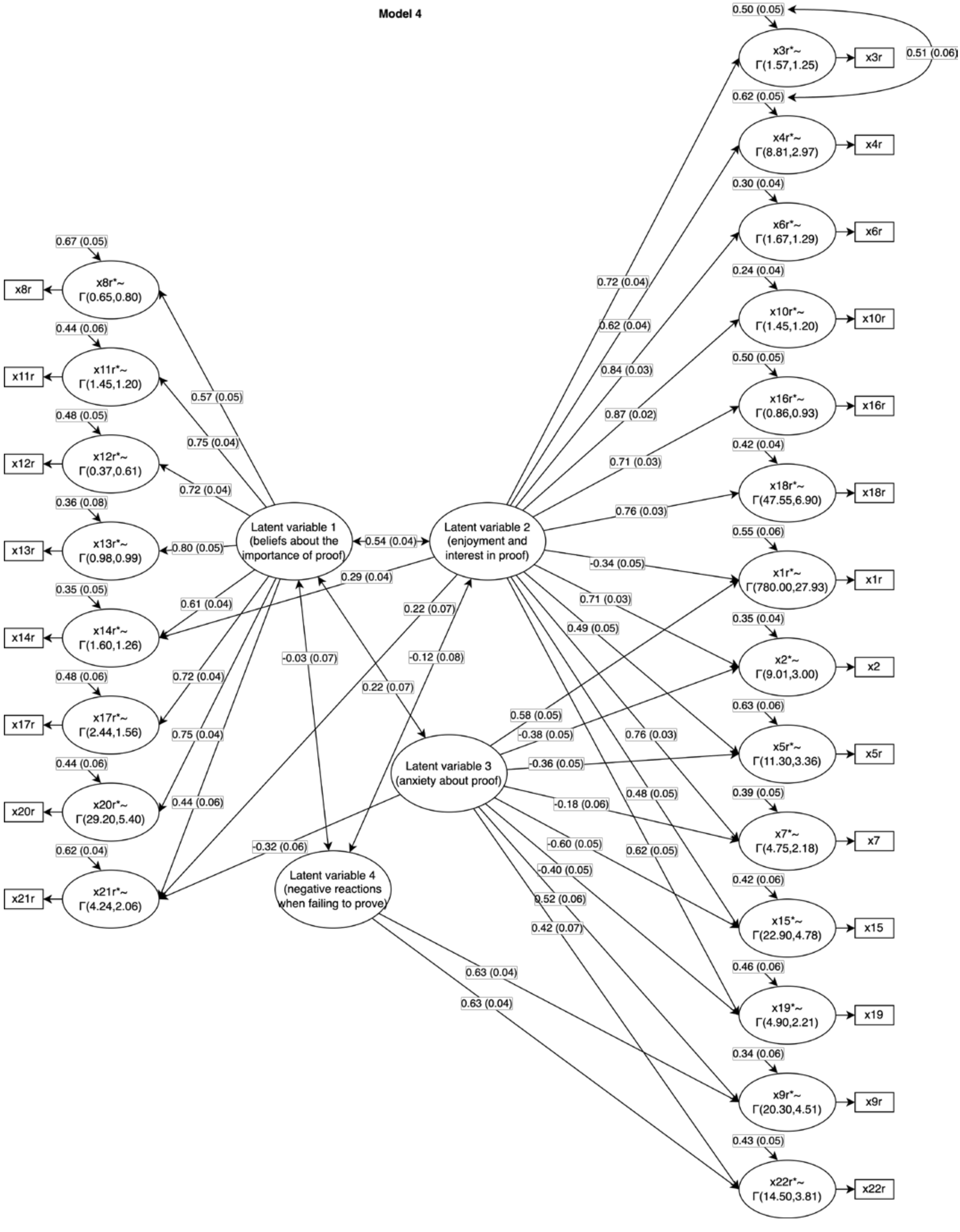


Figure 6. Model 4 (cat-LS-nonnorm) with Standardised Parameter Estimates (and Standard Errors)

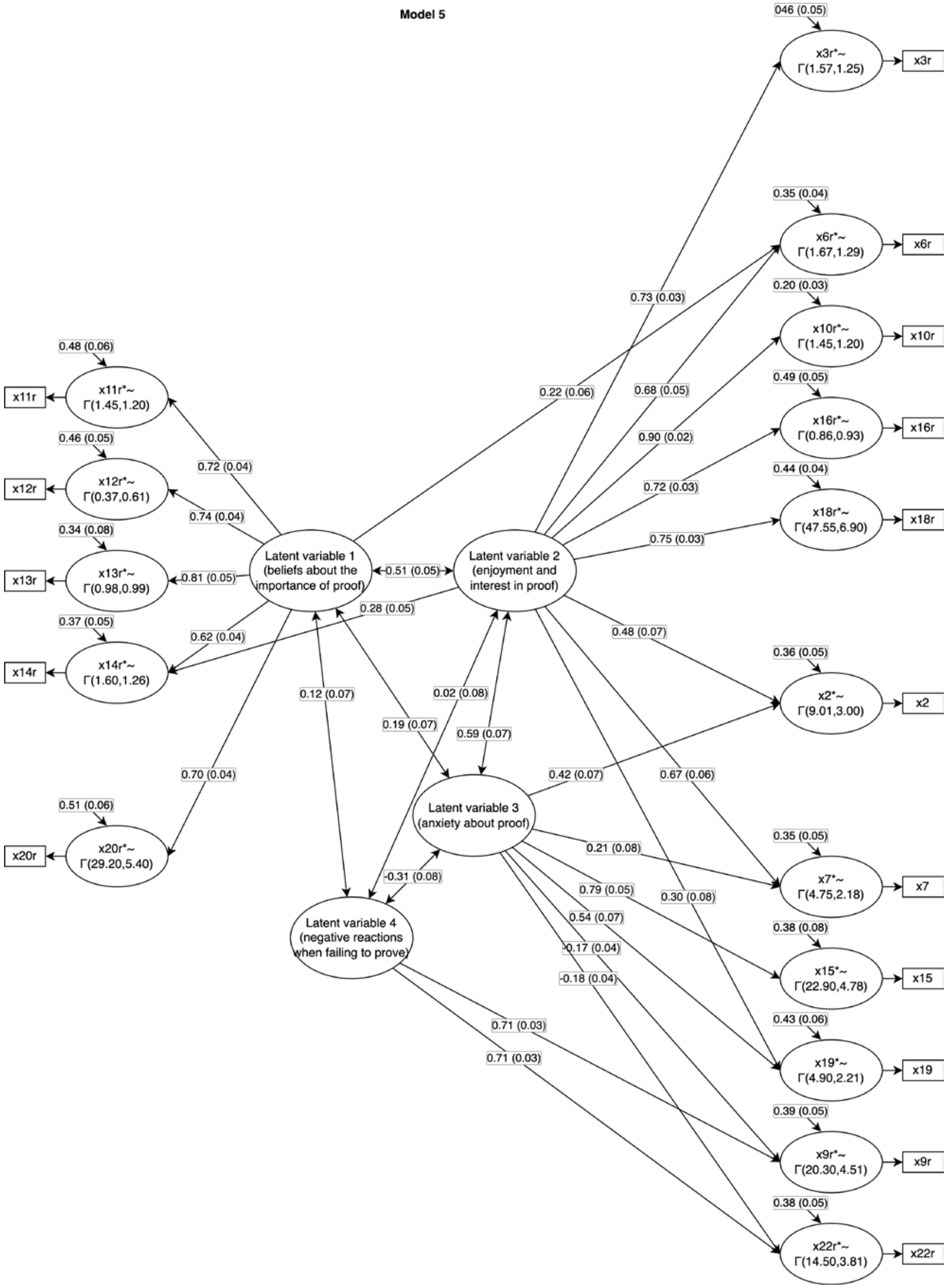


Figure 7. Model 5 (cat-LS-nonnorm) with Standardised Parameter Estimates (and Standard Errors)

Note that both cat-LS-norm and cat-LS-nonnorm resulted in Model 5 as their final model after the modification processes. However, they went through nuanced steps and numerical outputs. Model 5 had four latent variables, as same as Model 4, but with only 16 items used and a data-driven loading of x6 on 'importance' (Figure 7). Model 5 achieved 'good fit' under all cont-ML, cat-LS-norm and cat-LS-nonnorm (Table 4). Whilst all three conditions had consistent goodness-of-fit statistics, they had different results for some model parameters. In particular, the correlations between 'importance' and 'anxiety' (r_1) and between 'anxiety' and 'failing' (r_2) were statistically significant in cat-LS-norm (r_1=0.20, s.e.=0.08; r_2=-0.32, s.e.=0.09) and cat-LS-nonnorm (r_1=0.19, s.e.=0.07; r_2=-0.31, s.e.=0.08), but not in cont-ML (r_1=0.22, s.e.=0.10; r_2=-0.25, s.e.=0.35). In both Models 4 and 5, the results of modification indices were divided between cont-ML and cat-LS approaches. Together these showed that the choice of estimation method can impact modification of a model, and consequently, its final results and interpretation.

Discussion

This article has a goal of inviting further methodological discussions on SEM with ordinal variables in educational research. Recall that it is not my intention to argue which of cat-LS and cont-ML, or other methods, *should* be used in SEM, but this article aims to encourage researchers to clarify decisions on SEM. Doing so not only is good statistical practice (American Statistical Association, 2022; Wright, 2003), but also makes the analysis more transparent and accessible to both experienced and inexperienced readers (LeBeau et al., 2021). This can open up more potential engagement in SEM within educational research, and invite more reviews of research involving SEM. To the best of my knowledge, issues about debates on SEM with ordinal variables and its reporting have not yet been explicitly discussed in literature within educational research; this article reports the first review of practices of conducting and reporting SEM with ordinal variables in mathematics education research and proposes a standard for reporting. Compared to existing suggested practices, the standard highlights the necessity of reporting information about strategies for addressing nature of data, such as type of correlations and model estimation algorithm. If available, this information can enable reconstruction of analysis, inviting validation of results in a report.

The CFA example illustrates the proposed standard and contributes to research on differences between cat-LS and cont-ML (often simulation studies), as it demonstrates how different estimation methods can yield different results with real-life educational research data. The comparison of cat-LS and cont-ML can lead to questions relating to model identification, for example, *Which modified models should be retained if all are 'good fit'?* and *When is a model retained if results are inconsistent?* In the example, should Model 4 be rejected because it was rejected under cat-LS approaches? Did cont-ML yield overoptimistic results, accepting a poor model, or did cat-LS reject an acceptable model? In relation to the mathematics education context of the example, *to what extent, teachers' beliefs about the importance of proof are related to their anxiety about proof?* All three data treatment conditions yielded similar estimates of that parameter, but the difference in standard error has led to inconsistent significance test results. On the other hand, some results of the example between the approaches were consistent, and a question can be asked: *When, regarding research design, such as sample size, item scaling, data distribution, purposes and models tested, is it 'safe' to use cont-ML alone?*

As 'true' underlying distributions behind the ordinal data are unknown in real-life research, use of multiple data assumptions in CFA enables sensitivity analysis to examine robustness of the results. This will also address the above questions by identifying more confident results and interpretations. The proposed standard, illustrated by the reflection points and the CFA example, has been shown useful for strengthening research transparency, enabling reproduction of model parameters (see supplementary materials) and more detailed external reviews. Doing so, however, may increase research costs, such as expertise in and time costs for the analysis and its reporting. In addition, costs for expertise in reviewing SEM reports should also be considered; nonetheless, this article hopes to encourage peer review of SEM.

Moreover, this article only focuses on cat-LS and cont-ML. There exist other statistical innovations (e.g., Bayesian methods, item response theory) for examination of factor structure (for measurements) in research. Different methods have their own strengths and weaknesses. More research can be carried out to evaluate the results and costs of different methods, discussing their strengths and weaknesses. This will provide guidelines that balance research integrity and practicality in SEM and other statistical techniques.

Ethics Statements

The research study that underpins this publication received ethical approval from University of Oxford Central University Research Ethics Committee (Ref: ED-CIA-18-198) and The Education University of Hong Kong Human Research Ethics Committee (Ref: 2018-2019-0323). A Participant Information Sheet containing information about the study was attached to the survey as the cover sheet. Participants were asked to confirm that they had read the information sheet and were asked to sign on the sheet, confirming that they were willing to take part in the study.

Acknowledgements

The example presented in this article is based on the data of my doctoral study and dissertation (Lee, 2022). Part of this article had been presented at the 15th International Congress on Mathematical Education (7-14 July, 2024). I wish to thank people who have given valuable feedback on the draft of this article.

Conflict of Interest

The author has no relevant financial or non-financial interests to disclose.

Funding

Not applicable.

Generative AI Statement

Not applicable.

Data and Code Availability

The dataset that supports the example presented in this paper is available from Oxford Research Archive (ORA), which is available from the author upon request. The R code for all analyses reported in the article is available as supplementary materials.

References

Alabi, A. T., & Jelili, M. O. (2023). Clarifying likert scale misconceptions for improved application in urban studies. *Quality and Quantity*, *57*, 1337-1350. https://doi.org/10.1007/s11135-022-01415-8

Almeida, D. (2000). A survey of mathematics undergraduates' interaction with proof: Some implications for mathematics education. *International Journal of Mathematical Education in Science and Technology*, *31*(6), 869-890. https://doi.org/10.1080/00207390050203360

American Statistical Association. (2022). *Ethical guidelines for statistical practice*. American Statistical Association. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice

Andrews, P., & Diego-Mantecón, J. (2015). Instrument adaptation in cross-cultural studies of students' mathematics-related beliefs: Learning from healthcare research. *Compare: A Journal of Comparative and International Education*, *45*(4), 545-567. https://doi.org/10.1080/03057925.2014.884346

Brandenburg, N. (2024). Factor retention in ordered categorical variables: Benefits and costs of polychoric correlations in eigenvalue-based testing. *Behavior Research Methods*, *56*, 7241-7260. https://doi.org/10.3758/s13428-024-02417-0

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62-83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Cañete-Massé, C., Carbó-Carreté, M., Figueroa-Jiménez, M. D., Oviedo, G. R., Guerra-Balic, M., Javierre, C., Peró-Cebollero, M., & Guàrdia-Olmos, J. (2022). Confirmatory factor analysis with missing data in a small sample: Cognitive reserve in people with Down Syndrome. *Quality and Quantity*, *56*, 3363-3377. https://doi.org/10.1007/s11135-021-01264-x

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491. https://doi.org/10.1037/1082-989X.9.4.466

Foldnes, N., & Grønneberg, S. (2019). On identification and non-normal simulation in ordinal covariance and item response models. *Psychometrika*, *84*(4), 1000-1017. https://doi.org/10.1007/s11336-019-09688-z

Foldnes, N., & Grønneberg, S. (2020). Pernicious polychorics: The impact and detection of underlying non-normality. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 525-543. https://doi.org/10.1080/10705511.2019.1673168

Foldnes, N., & Grønneberg, S. (2022). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. *Psychological Methods*, *27*(4), 541-567. https://doi.org/10.1037/met0000385

Frasier, B. J. (2010). *Secondary school mathematics teachers' conceptions of proof* (Publication No. 3417066) [Doctoral dissertation, University of Massachusetts Lowell]. ProQuest Dissertations & Theses Global.

Garson, G. D. (2015). *Structural equation modeling*. Statistical Associates Publishing.

Geisler, S., Rolka, K., & Rach, S. (2023). Development of affect at the transition to university mathematics and its relation to dropout - identifying related learning situations and deriving possible support measures. *Educational Studies in Mathematics*, *113*, 35-56. https://doi.org/10.1007/s10649-022-10200-1

Grønneberg, S., & Foldnes, N. (2024). Factor analyzing ordinal items requires substantive knowledge of response marginals. *Psychological Methods, 29*(1), 65-87. https://doi.org/10.1037/met0000495

Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial least squares structural equation modeling (PLS-SEM) using R*. Springer. https://doi.org/10.1007/978-3-030-80519-7

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity*, *44*, 153-166. https://doi.org/10.1007/s11135-008-9190-y

Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6-23. https://doi.org/10.1037/a0014694

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443-482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G. (1994). Structural equation modeling with ordinal variables. *Multivariate Analysis and Its Applications: IMS Lecture Notes - Monograph Series*, *24*, 297-310. https://doi.org/10.1214/lnms/1215463803

Keçeli-Bozdağ, S., Uğurel, I., & Bukova-Güzel, E. (2015). Development of attitude scale towards proof and proving: The case of mathematics student teachers. *Kastamonu Education Journal/Kastamonu Eğitim Dergisi*, *23*(4), 1585-1600. http://bit.ly/3IVbbcl

Khine, M. S. (Ed.). (2013). *Application of structural equation modeling in educational research and practice*. Sense Publishers. https://doi.org/10.1007/978-94-6209-332-4.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Knuth, E. J. (2002). Secondary school mathematics teachers' conceptions of proof. *Journal for Research in Mathematics Education*, *33*(5), 379-405. https://doi.org/10.2307/4149959

Kolbe, L., Oort, F., & Jak, S. (2021). Bivariate distributions underlying responses to ordinal variables. *Psych*, *3*(4), 562-578. https://doi.org/10.3390/psych3040037

Kotelawala, U. M. (2007). *Exploring teachers' attitudes and beliefs about proving in the mathematics classroom* (Publication Number 3266716) [Doctoral dissertation, Columbia University]. ProQuest Dissertations & Theses Global.

LeBeau, B., Ellison, S., & Aloe, A. M. (2021). Reproducible analyses in education research. *Review of Research in Education*, *45*(1), 195-222. https://doi.org/10.3102/0091732X20985076

Lee, G. C.-Y. (2022). *Hong Kong preservice teachers' beliefs and attitudes towards teaching proof in school mathematics: A design-based research* [Doctoral dissertation, University of Oxford]. Oxford University Research Archive. http://bit.ly/45ieKB5

Lenz, K., Reinhold, F., & Wittmann, G. (2024). Topic specificity of students' conceptual and procedual fraction knowledge and its impact on errors. *Research in Mathematics Education, 26*(1), 45-69. https://doi.org/10.1080/14794802.2022.2135132

Li, C.-H. (2016a). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936-949. https://doi.org/10.3758/s13428-015-0619-7

Li, C.-H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, *21*(3), 369-387. https://doi.org/10.1037/MET0000093

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1-30. https://doi.org/10.18637/jss.v085.i04

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132. https://doi.org/10.1007/BF02294210

Narayanan, A. (2012). A review of eight software packages for structural equation modeling. *Statistical Computing Software Reviews*, *66*(2), 129-138. https://doi.org/10.1080/00031305.2012.708641

Nyaumwe, L., & Buzuzi, G. (2007). Teachers' attitudes towards proof of mathematical results in the secondary school curriculum: The case of Zimbabwe. *Mathematics Education Research Journal*, *19*, 21-32. https://doi.org/10.1007/BF03217460

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 557-595. https://doi.org/10.1207/S15328007SEM0704_3

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2023). R: A language and environment for statistical computing (Version 4.3.1) [Computer Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373. https://doi.org/10.1037/a0029315

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5*, Article 589965. https://doi.org/10.3389/feduc.2020.589965

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 149-160. https://doi.org/10.1080/10705511.2013.824793

Savalei, V. (2021). Improving fit indices in structural equation modeling with categorical data. *Multivariate Behavioral Research*, *56*(3), 390-407. https://doi.org/10.1080/00273171.2020.1717922

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201-223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-338. https://doi.org/10.3200/JOER.99.6.323-338

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Psychology Press. https://doi.org/10.4324/9781410610904

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, *2*, 1-19. https://doi.org/10.1146/annurev-statistics-010814-020127

Street, K. E. S., Malmberg, L.-E., & Stylianides, G. J. (2022). Changes in students' self-efficacy when learning a new topic in mathematics: A micro-longitudinal study. *Educational Studies in Mathematics*, *111*, 515-541. https://doi.org/10.1007/s10649-022-10165-1

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*(1), 123-136. https://doi.org/10.1348/000709903762869950

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

Zengin, Y. (2017). The effects of GeoGebra software on pre-service mathematics teachers' attitudes and views towards proof and proving. *International Journal of Mathematical Education in Science and Technology*, *48*(7), 1002-1022. https://doi.org/10.1080/0020739X.2017.1298855

Zhang, Y., Yang, X., Sun, X., & Kaiser, G. (2023). The reciprocal relationship among Chinese senior secondary students' intrinsic and extrinsic motivation and cognitive engagement in learning mathematics: A three-wave longitudinal study. *ZDM Mathematics Education*, *55*, 399-412. https://doi.org/10.1007/s11858-022-01465-0

Appendix

This appendix summarises the proposed standard as a checklist. When preparing for a SEM analysis, researchers should know their hypothesised model(s), know their data, and reflect on information about the analysis to be included in a research report.

Know the hypothesised models

1. Is foundation of hypothesised model(s) sound?
2. Are the instrument(s) selected suitable for the models?

Know the data

1. What type of data are collected for observed variables?
2. Are the data considered to be ordered-categorical (ordinal), interval or continuous?
3. [If considered to be interval or continuous] does the data deviate from normality?
4. [If considered to be ordinal] what latent distribution is the data assumed?
5. How are missing values addressed?
6. What type(s) of correlations (e.g., Pearson, polychoric) are the data suitable for?
7. Is the data (or correlation/covariance matrix) available?

Reflect on the analysis methods

1. What software is used for analysis?
2. What model estimation algorithm(s) are used in the analysis?
3. Are adjustment(s) applied to model parameter estimates, standard errors and/or model fit statistics?
4. How are the model(s) evaluated?
    a. What are criteria for model acceptance or rejection or comparison?
    b. What are fit statistics of the model(s)?
5. Is model modification involved?
6. Is software code available?

Reflect on the reporting

Does the reporting enable readers to re-construct the analysis?

[i] Absolute agreement was computed by dividing the number of ratings (Yes/No) for a code that were the same between the first and last coding time points by the total number of ratings for the code.