Original Research Article

Educational and Psychological Measurement
2023, Vol. 83(5) 885–906

DOI: 10.1177/00131644221130482
journals.sagepub.com/home/epm
SAGE

Equidistant Response Options on Likert-Type Instruments: Testing the Interval Scaling Assumption Using Mplus

Georgios Sideridis[1,2], Ioannis Tsaousis[2] and Hanan Ghamdi[3]

Abstract

The purpose of the present study was to provide the means to evaluate the "interval-scaling" assumption that governs the use of parametric statistics and continuous data estimators in self-report instruments that utilize Likert-type scaling. Using simulated and real data, the methodology to test for this important assumption is evaluated using the popular software Mplus 8.8. Evidence on meeting the assumption is provided using the Wald test and the equidistant index. It is suggested that routine evaluations of self-report instruments engage the present methodology so that the most appropriate estimator will be implemented when testing the construct validity of self-report instruments.

Keywords

Likert-type, interval scaling, survey assumptions, Mplus

Since the pioneering work of Guttman, Thurstone, and Likert, several attempts to categorize the continuum of traits, skills, and abilities have been introduced (see Nunnally, 1978). Among the most prevalent are Likert-type scales (Likert, 1932) that

[1]Harvard Medical School, Boston, MA, USA
[2]National and Kapodistrian University of Athens, Greece
[3]Education and Training Evaluation Commission, Riyadh, Saudi Arabia

Corresponding Author:
Georgios Sideridis, Harvard Medical School, Boston, MA 02115, USA.
Email: georgios.sideridis@gmail.com

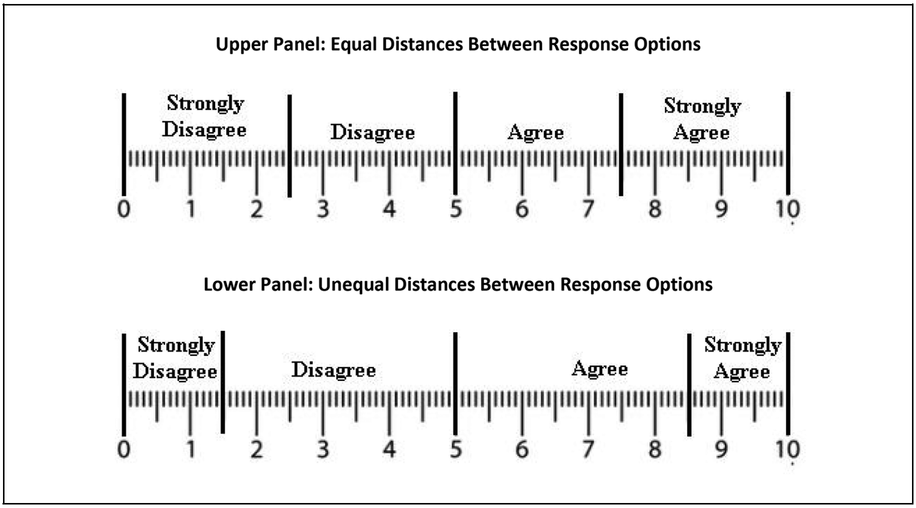

Figure 1. Assuming Equal Interval Scaling (Upper Panel) or Unequal (Lower Panel).
Note. The unequal threshold system suggests that the frequency of responses lessens as they become more extreme (strong disagreements or agreements). Upper panel: Equal distances between response options. Lower panel: Unequal distances between response options.

usually engage a disagreement-to-agreement ordering. Thus, Likert-type data represent a categorical representation of a latent unidimensional continuous construct (Andrich, 1978b; Spratto, 2018). The present study deals with the inherent researcher-made assumption underlying these scales that adjacent options in ordered categorical schemes are equidistant (Henkel, 1975). In other words, the distance between strongly disagree and disagree is the same as the one between agree and strongly agree. Certainly, this is tentative and, should for that reason be, a testable assumption (Hensler & Stipak, 1979).

Figure 1 shows an equal and unequal spacing scheme using Likert-type disagree-to-agree scaling. As shown in the upper panel of Figure 1, the distance between adjacent disagree and agree options is fixed to 2.5 raw units of a ruler. This figure suggests that the conceptual distance between *disagree* and *strongly disagree* is the same as the one between *agree* and *strongly agree*. If, however, one considers that strong disagreements/agreements are positioned at the tails of a normally distributed latent variable describing the disagreement-to-agreement continuum, then one may favor the unequal spacing system shown in the lower panel of Figure 1, which suggests that more extreme responses (i.e., ''strong'' feelings) are less likely to occur compared with less extreme responses, thus, occupying less measurement space. If the above example is extended to the inclusion of a 6-point Likert-type scaling in which responses are *absolute disagreement, major disagreement, minor disagreement, neither agree nor disagree, minor agreement, major agreement, and complete agreement*, it is even harder to conceive equal conceptual spacing between response options as there aren't any good reasons to suggest that the distance between *absolute and major disagreement* is the same as the one between *major disagreement and minor disagreement*. The latter likely represents a larger conceptual distance compared with the former.

The necessity of verifying the ''equal interval'' assumption of Likert-type scaling systems stems from the unjustified use of parametric statistical methods using continuous estimators in the analyses of these data as they are oftentimes treated as being continuous. Based on the pioneering work of Stevens (1946), a salient difference between ordinal and interval measurement is that the latter assumes both rank ordering but also equal spacing. If the data do not satisfy equal interval scaling, parametric techniques assuming normality and interval measurement are invalid, resulting in the distortion of model fit statistics (Hutchinson & Olmos, 1998; Vigderhous, 1977). There is a solution to this potential problem: there are several analytical models that do not require the interval scaling assumption such as the graded response model (GRM, Muraki, 1992) or Andrich's (1978a) model for polytomous responses. What is problematic is if the empirical study literature points to the misuse of proper analytical methodologies with categorical data when they are treated as being continuous (see Martens, 2005).

Here, an important distinction needs to be made which refers to the types of models employed. Using measured variables, tests of point estimates such as *t*-tests and ANOVA have shown robustness in assuming interval scaling (Gaito, 1980; Garifio & Perla, 2007). However, results from simulation studies using latent variable modeling have shown that modeling Likert-type data as continuous has been associated with biased parameter estimates as the continuous factor model is essentially misspecified when used with ordinal data (Beauducel & Herzberg, 2006; Rhemtulla et al., 2012), especially when the number of response options is small (Johnson & Creech, 1983. There is unequivocal evidence that asymmetric category thresholds bias the maximum likelihood estimator when being treated as symmetric (Babakus et al., 1987; DiStefano, 2002; Dolan, 1994; Forero et al., 2009; Green et al., 1997). For example, the simulation study by Beauducel and Herzberg (2006) showed that under all conditions of the confirmatory factor analysis (CFA) model (1 factor model with 5 items to 8 factor model with 40 items) number of items, number of categories (2–6) and sample sizes (250–1,000) the maximum likelihood (ML) estimator was inferior compared with the diagonally weighted least squares estimator. Even the robust estimators of ML (Satorra & Bentler, 1988, 1994; Yuan & Bentler, 1998; Yuan et al., 2011) showed biases in the presence of asymmetrical thresholds (Potthast, 1993; Rigdon & Fergusson, 1991). As the number of options increases to five or more, the bias becomes smaller as the ordered variables approach continuity (Johnson & Creech, 1983; B. O. Muthen & Kaplan, 1985; Rhemtulla et al., 2012; Sullivan & Artino, 2013; Zumbo & Zimmerman, 1993), although these studies have not focused on the distances across response options. This conclusion is in line with the observation of Bollen and Barb (1981) that a scaling system with 1–4 options will be weakly related to a latent construct compared with being measured in a continuous manner. Others have pointed to the inappropriateness of assuming rather than confirming the interval scaling assumption (Jakobsson, 2004; Jamieson, 2004; Knapp, 1990; O'Brien, 1979; Pohl, 1981; Spector, 1980). To evaluate whether validity studies ''err'' in that direction a systematic review of the literature was conducted in major counseling psychology journals as our present evaluation dealt with a well-known instrument in health and well-being, the Positive Youth Development Inventory (PYDS, Arnold et al., 2012). The goal of this review was to evaluate the extent to which studies ascertain construct validity using analytical means that are appropriate for continuous data, without testing whether ordered data satisfy interval scaling measurement. The present evaluation comes following early concerns that structural equation modeling (SEM) practices observed in counseling psychology journals do not engage in best practices as recommended by SEM experts (Martens, 2005). These results are briefly described in the next section.

A Review of Validity Studies in Counseling Psychology

A systematic review of the literature on the use of appropriate analytical methods with ordered data was conducted. Fifty-seven validity studies published between 2018 and 2021 from the journals: *Measurement and Evaluation in Counseling and Development, Journal of Counseling and Development, Journal of Counseling Psychology, and Professional School Counseling* met the criteria for inclusion in this brief review. The extraction procedure involved a person reading all titles, abstracts, and, subsequently, the manuscripts to verify appropriateness for inclusion and availability of pertinent data. Specific inclusionary criteria involved: (a) validation of an instrument, (b) use of a psychometrics model such as exploratory factor analysis, CFA, or item response theory (IRT), (c) use of a categorical/ordered scaling system, and (d) inclusion of information about the analytical methodology employed. Appropriate analytical methods involved IRT and CFA methodologies and estimators that accounted for the categorical nature of the data (e.g., maximum likelihood robust, weighted least squares mean and variance adjusted) or non-normality (e.g., Bayesian methodologies). As shown in Table 1, 33 of the published studies (57.9%) specifically mentioned methodologies that were appropriate for categorical/ordered data. Interestingly, 24 studies (42.1%) either did not mention or ignored the categorical nature of the data by assuming that data were continuous, meeting the assumption of equidistant measurement. Of these, 42.1% there is likely a large proportion that utilized inappropriate analytical methods. This raises serious concerns about the implicit assumption that ordered data (such as those utilizing Likert-type scales) can be treated as continuous, in that they meet the interval scaling assumption inherent in continuous data.

The purpose of the present study was to evaluate the equal interval assumption when validating instruments using Likert-type scaling systems. Using the popular software Mplus 8.8 (Muthén & Muthén, 1998–2018) routines for evaluating interval scaling at the item level are provided in Appendix A. For that purpose, the GRM

Table 1. Description of Validity Studies, Scaling, and Analytical Methodologies.

Study	Data scaling	Analytical method	Estimator	Appropriate for categorical data
Akdogan et al. (2018)	Likert	CFA	N.A.	No
Aldawsari et al. (2021)	Likert	CFA	N.A.	No
Autin et al. (2019)	Likert	CFA	N.A.	No
Bardoshi et al. (2019)	Likert	CFA	N.A.	No
Blau & DiMino (2019)	Likert	CFA	N.A.	No
Bloom & Dillman-Taylor (2020)	Likert	EFA	N.A.	No
Cho et al. (2018)	Likert	CFA	MLMV	Yes
Cimsir & Akdogan (2020)	Likert	CFA	MLR	Yes
Dillman-Taylor et al. (2019)	Likert	CFA	MLR	Yes
Erford et al. (2021)	Likert	CFA	WLSMV	Yes
Fu & Zhang (2019)	Likert	CFA	ML	No
Ganho-Avila et al. (2019)	Likert	CFA	MLR	Yes
Ghabrial & Andersen (2020)	Likert	CFA	ML	No
Ghosh et al. (2021)	Likert	CFA	WLSMV	Yes
Gonzalez et al. (2021)	Likert	CFA	N.A.	No
Greene (2019)	Likert	EFA	N.A.	No
Griffin et al. (2018)	Likert	CFA	MLR	Yes
Halamova et al. (2021)	Likert	IRT	MLR	Yes
Hiles-Howard et al. (2019)	Likert	CFA	ML	No
Johnson & Karcher (2019)	Likert	CFA	FIML	No
Johnson et al. (2021)	Likert	CFA	FIML	No
Kim et al. (2021)	Likert	CFA	FIML	No
Kivlighan et al. (2018)	Likert	CFA	Bayes	Yes
Lau et al. (2019)	Likert	CFA	ML	No
Lee et al. (2019)	Likert	CFA	N.A.	No
Lee et al. (2021)	Likert	CFA	WLSMV	Yes
Levant & Parent (2019)	Likert	IRT	N.A.	Yes
Levant et al. (2020)	Likert	CFA	MLR	Yes
Lim & Kim (2020)	Likert	CFA	WLSMV	Yes
Liu et al. (2018)	Likert	CFA	WLSMV	Yes
Lu et al. (2018)	Likert	CFA	Bollen-Stine	Yes
Ludlow et al. (2019)	Likert	IRT	JMLE	Yes
Luo et al. (2021)	Likert	CFA	ML	No
Martin et al. (2020)	Likert	CFA	WLSMV	Yes
Mazahreh et al. (2019)	Likert	CFA	Bollen-Stine	Yes
Moate et al. (2019)	Likert	CFA	MLR	Yes
Oh & Shillingofrd-Butler (2021)	Likert	CFA	MLR	Yes
Pederson et al. (2021)	Likert	CFA	N.A.	No
Perez-Rojas et al. (2019)	Likert	IRT	N.A.	Yes
Poynton et al. (2019)	Likert	IRT	N.A.	Yes
Pozza et al. (2019)	Likert	CFA	N.A.	No
Rowan-Kenyon et al. (2021)	Likert	CFA	Bollen-Stine	Yes
Shea et al. (2019)	Likert	CFA	MLR	Yes
Shin et al. (2018)	Likert	CFA	MLR	Yes
Simons (2018)	Likert	EFA	N.A.	No

(continued)

Table 1. (continued)

Study	Data scaling	Analytical method	Estimator	Appropriate for categorical data
Swank et al. (2020)	Nominal	CFA	WLSMV	Yes
Tadlock-Marlo & Hill (2019)	Likert	IRT	N.A.	Yes
TaeHyuk-Keum et al. (2018)	Likert	CFA	MLR	Yes
Toland et al. (2021)	Likert	IRT	ML-EAP	Yes
Trub et al. (2020)	Likert	CFA	FIML	No
Veronese & Pepe (2019)	Likert	CFA	N.A.	No
Waldrop et al. (2019)	Likert	CFA	WLSMV	Yes
Wang et al. (2019)	Likert	CFA	ML	No
Watson et al. (2019)	Likert	CFA	Bollen-stine	Yes
Watson et al. (2020)	Likert	CFA	ML	No
Xavier et al. (2019)	Likert	CFA	WLSMV	Yes
Young & Bryan (2018)	Likert	CFA	Bollen-Stine	Yes

Note. CFA = confirmatory factor analysis; N.A. = not available; EAP = Expected A posteriori; EFA = exploratory factor analysis; MLMV = maximum likelihood mean and variance adjusted; MLR = maximum likelihood robust; WLSMV = weighted least squares mean and variance adjusted; ML = maximum likelihood; IRT = item response theory; FIML = full information maximum likelihood; JMLE = joint maximum likelihood estimation.

(Muraki, 1992) was utilized using the measurement of *caring* from the PYDS (Arnold et al., 2012).

Method

Participants

Data came from a random sample of 500 participants (from a total of 5,443) who took the PYDS as part of participating in online assessments via a governmental platform in the Kingdom of Saudi Arabia. Our choice to select a random sample was based on avoiding excessive levels of power and observing trivial, albeit significant effects. Using the MacCallum et al. (1996) approach, the power for a unidimensional CFA model with 8 indicators was 85% when contrasting an acceptable (i.e., RMSEA=0.05) from an unacceptable model (i.e., RMSEA=0.08). Furthermore, a Monte Carlo simulation was run to ensure the stability of the slopes and intercepts as per a 2PL model. Results indicated that using 500 replicated samples with n = 500 power levels were equal to 99% for discrimination parameters equal to 1 and intercept terms equal to 0.5. Coverage for these parameters ranged between 94.2% and 96.6%. Thus, our proposed sample size would likely result in unbiased estimates of model parameters without the burden of enhanced power levels.

Instrument. Arnold et al. (2012) developed the PYDS. The PYDS utilizes 55 items to assess six basic attributes, namely, Competence (14 items), Character (9 items), Connection (8 items), Caring (8 items), Confidence (9 items), and Contribution (7 items). Participants respond using a 4-point Likert-type scale anchored from 1 = *strongly disagree* to 4 = *strongly agree*. Using translation and back translation methodologies (Brislin, 1970), the original scale was translated into Arabic. To control for response bias (i.e., extreme response style, acquiescence, see Bolt & Johnson, 2009), developers of the Arabic version chose to reverse-code eight items. Validity studies have confirmed the factorial structure of the Arabic PYDS (Tsaousis et al., 2021). For the purposes of the present evaluation, the 8-item *Caring* domain was used that assesses a person's proclivity to be empathic and caring toward others (see Appendix B). It consists of eight items that utilize a 4-point scaling system anchored between strongly disagree and strongly agree. The mid-response option that was available in the original version was deleted from the Arabic version of the scale in support of several criticisms raised about the necessity of the middle option and concerns about the midpoint response style.

Data Analysis: GRM to Test Simple Structures With Ordered Data

Given the polytomous nature of the data, the GRM (Muraki, 1992; Samejima, 1969) was deemed the appropriate choice for these data. Among polytomous models, there are two main categories: The different models that involve cumulative probabilities and the adjacent models. Adjacent models estimate the probability of moving from one response option to the next. However, because they are amenable to reversals and can result in the measurement of unstable parameters, especially with small samples (e.g., Spratto, 2018), they were not utilized here. Instead, we used the cumulative model in which the probability of responding to one category is contrasted against the sum of the options above it. For example, the probability of responding to Category 1 (e.g., *strongly disagree*) is contrasted with the probability of selecting any one of the categories above Category 1 (i.e., 2, 3, 4, or *strongly disagree* against the options *disagree–agree–strongly agree*).

The GRM represents an extension of the typical 2PL model, which has a single discrimination parameter (α_i) and a single item location (β_i). The extension involves the presence of thresholds equal to the number of categories–1 (termed $\beta_{i\kappa}$) and, thus, operating characteristic curves (OCCs) are fit for each threshold representing the boundary between two adjacent categories. The conditional probability of item endorsement is estimated as follows:

$$P\left(Y_{ij}=1|a_i, b_{i\kappa}\right)=\frac{e^{a_i(\theta_j-\beta_{i\kappa})}}{1+e^{a_i(\theta_j-\beta_{i\kappa})}}, \quad (1)$$

Thus, the probability that person *j* will endorse the threshold (β_κ) of item *i* is a function of a single item discrimination parameter α_i and person's *j* ability level theta (θ_j).

A visual examination of a sample item (Item 5 from the Caring subscale of the PYDS-see all items in Appendix B) is shown in Figures 2 and 3 to illustrate the merits of the cumulative logit model. Figure 2 displays an OCC and Figure 3 a category

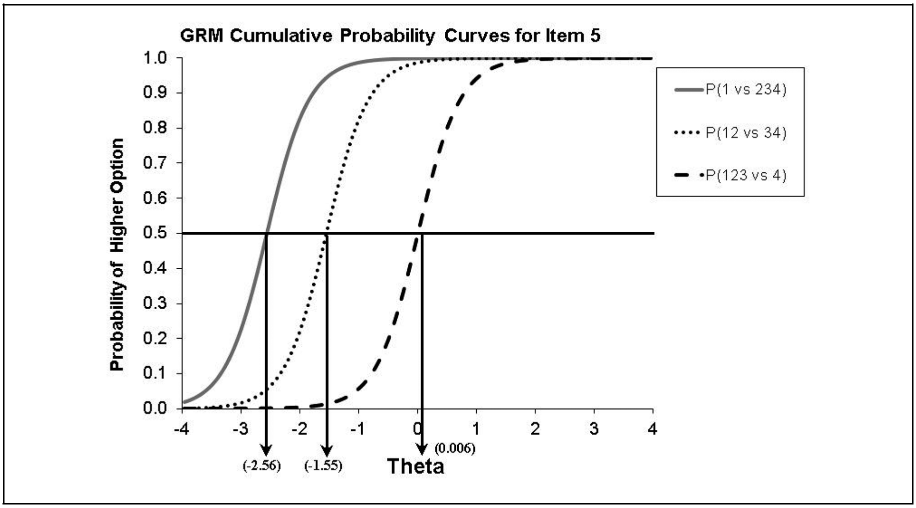

Figure 2. OCCs for Item 5 of the "Caring" Subscale of the PYDS With Category Boundary Locations $b_{1\text{-}i5} = -2.56$, $b_{2\text{-}i5} = -1.55$, and $b_{3\text{-}i5} = 0.006$.

Note. The vertical lines point to the theta estimate required by a person to have a 50% chance of responding in that category or higher. GRM = graded response model; OCCs = operating characteristic curves; PYDS = Positive Youth Development Scale.

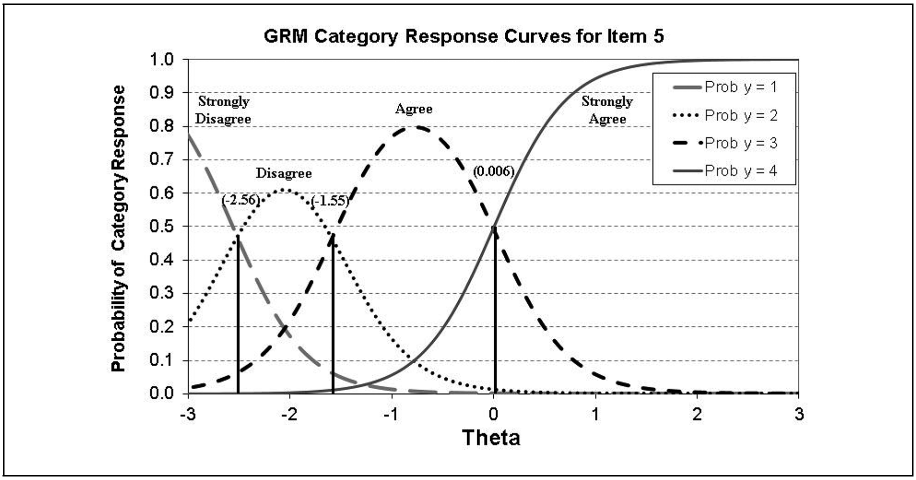

Figure 3. CCCs for Item 5 of the "Caring" Subscale of the PYDS (i.e., Other People's Feelings Matter to Me).

Note. CCCs = category characteristic curves; GRM = graded response model; PYDS = Positive Youth Development Scale; T1 = theta required to respond disagree compared with strongly disagree; T2 = theta required to respond agree compared with disagree; T3 = theta required to respond strongly agree compared with agree.

characteristic curve for Item 5 of the caring subscale. As shown in Figure 2, the first threshold was –2.56 suggesting that individuals with theta levels equal to –2.56 and below had a 50% chance to endorse *strongly disagree*. In other words, individuals with very low caring levels would strongly disagree with the statement that ‘‘Other people's feelings matter to me.'' Caring levels (theta) would have to be increased to –1.55 for individuals to have a 50% chance to select the *disagree* options (1 and 2), compared with the two *agree* options (3 and 4). Last, above zero thetas (i.e., θ =0.006) would be required for a person to select *strongly agree* compared with the three previous options.

Two analytical means were involved to assess the equidistance of the response options: the Wald test, and the equidistance index. The Wald test was employed to evaluate the equidistance hypothesis using a stepwise approach (testing one item at a time). The equidistance index proposed by Spratto (2018) was used as the difference of differences that is subtracting the difference between thresholds 1 versus 2 from 2 versus 3. A value of the index equal to zero would suggest equidistance among response options and satisfaction with the interval scaling assumption. Positive values of the index would suggest that the distance between Thresholds 2 and 3 is greater compared with the distance between Thresholds 1 and 2 and vice versa. Since the metric of the index is in logits, it is easy to understand the magnitude of non-equivalent thresholds but also to evaluate them using effect size indicators as suggested earlier (e.g., Dorans & Holland, 1992). For example, threshold differences within an item were termed small when <0.45 logits, medium when between 0.45 and 0.89 logits and large when $\geq$0.90 logits (see also Holland & Thayer, 1988; Zwick et al., 1999).

Testing the Equivalence of Thresholds: An Example Using the Construct of "Caring"

After fitting the GRM model to the data results indicated acceptable model fit (–2LL=6,703.68; M_2[244]=556.33, p<.001, RMSEA=0.05; AIC=6,767.68; BIC=6,902.55). Marginal reliability was equal to 0.81, which was acceptable. Figure 4 shows the Test Information Function in which the area of largest sensitivity is to the left of the figure (suggesting that the measure is highly sensitive to low levels of the latent trait).

After specifying equidistant thresholds, results indicated that all but two items failed to exhibit equal intervals (see Table 2) using the Wald test. The equidistant items were Numbers 4 and 8. Given that only 2/8 items (25%) showed equidistance in thresholds, the treatment of the current ordered data as continuous is prohibited. Further information was provided by the equidistance index. After subtracting 1–2 difference in thresholds from the 2–3 difference, results indicated that the distance between Thresholds 2 and 3 was larger compared with that of Thresholds 1 and 2, and this finding was consistent across all items but varied in magnitude. As shown in Table 2, there were three effects larger than large, three medium sized ones, and two

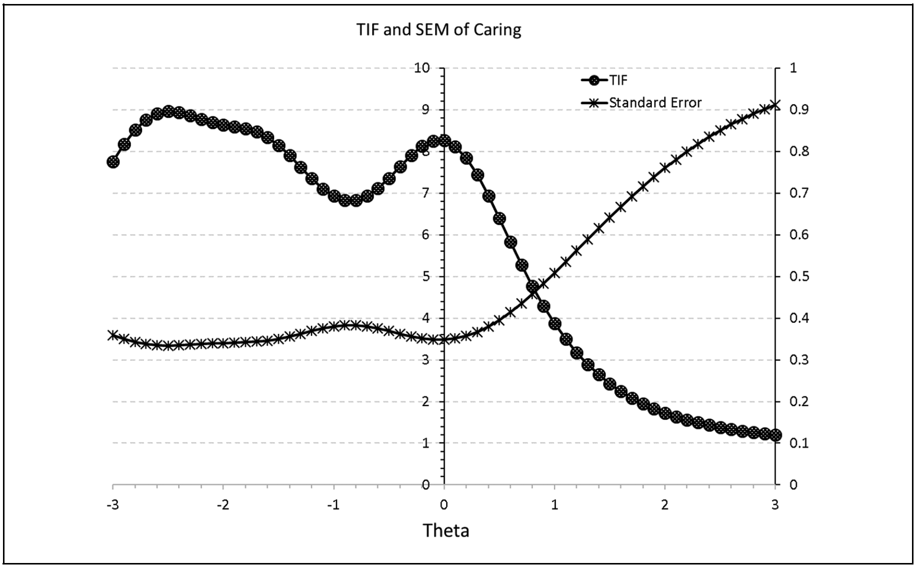

Figure 4. TIF and Conditional SEM of the Caring Subscale of the PYDS.
Note. The scaling at zero is for information and the secondary scaling is at +3 logits for SEM. PYDS = Positive Youth Development Scale; TIF = Test Information Function; SEM = Standard Error of Measurement.

that fall in the ''small'' range. This finding suggests that given the present sample and the current scale the distance between strongly disagree and disagree was smaller compared with the conceptual distance between the options agree and strongly agree. Thus, with the present Likert-type data, the assumption of interval scaling was violated. Consequently, in principle, the use of estimators in factor analytic models that do not account for the categorical nature of the data will be inappropriate and would hence likely lead to erroneous conclusions. Furthermore, the employment of estimators that account for non-normal, non-equidistant data needs to be utilized.

Concluding Remarks

The present study addresses a controversial issue in measurement, that of assuming that ordered data as Likert-type are, likely to possess the properties of continuous data. Among assumptions, that of interval scaling is one of the most prevailing as it defines continuous data. A systematic review of the literature including studies from the last 3 years in counseling psychology suggested that $>40\%$ of the studies engage parametric statistical models and assume that Likert-type data possess the properties of continuous data (particularly that of interval scaling). An analysis with real data from the construct of the caring subscale of the PYDS suggested that response categories were not equidistant and, consequently, data should not be analyzed using

Table 2. Equidistance Between Thresholds in the Caring Subscale of the PYDS.

Item	Discrimination	Threshold/	*SE*	Diff. thresholds/	*SE* of diff.	E.S. convention	*Z*-test	Wald test	*df*	Equidistance index
Item 1	1.866	–3.306	0.338	—	—	—	—	—	—	—
Threshold 1–2		–2.825	0.266	0.481	0.210	—	—	—	—	—
Threshold 2–3		–0.488	0.080	2.338	0.249	—	—	—	—	—
Diff-Thr. Item 1		—	—	–1.857	0.327	Large	–5.678***	32.241***	1	1.857
Item 2	1.992	–2.911	0.268	—	—	—	—	—	—	—
Threshold 1–2		–1.683	0.136	1.228	0.212	—	—	—	—	—
Threshold 2–3		0.098	0.074	1.781	0.143	—	—	—	—	—
Diff-Thr. Item 2		—	—	–0.554	0.233	Medium	–2.381*	5.670*	1	0.554
Item 3	1.061	–4.550	0.568	—	—	—	—	—	—	—
Threshold 1–2		–2.057	0.230	2.493	0.434	—	—	—	—	—
Threshold 2–3		0.770	0.129	2.828	0.295	—	—	—	—	—
Diff-Thr. Item 3		—	—	–0.335	0.398	Small	–0.841	0.708	1	0.335
Item 4	1.490	–3.324	0.344	—	—	—	—	—	—	—
Threshold 1–2		–2.288	0.218	1.036	0.227	—	—	—	—	—
Threshold 2–3		–0.266	0.085	2.021	0.207	—				—
Diff-Thr. Item 4		—	—	–0.986	0.277	Large	–3.564***	12.701***	1	0.985
Item 5	2.797	–2.563	0.212	—	—	—	—	—	—	—
Threshold 1–2		–1.552	0.112	1.011	0.174	—	—	—	—	—
Threshold 2–3		0.062	0.065	1.614	0.116	—	—	—	—	—
Diff-Thr. Item 5		—	—	–0.602	0.200	Medium	–3.009**	9.054*	1	0.602
Item 6	1.479	–3.421	0.354	—	—	—	—	—	—	—
Threshold 1–2		–1.999	0.188	1.421	0.260	—	—	—	—	—
Threshold 2–3		0.088	0.085	2.087	0.196	—	—	—	—	—
Diff-Thr. Item 6		—	—	–0.666	0.284	Medium	–2.345*	5.501*	1	0.666
Item 7	2.530	–2.609	0.226	—	—	—	—	—	—	—
Threshold 1–2		–1.804	0.135	0.806	0.170	—	—	—	—	—
Threshold 2–3		–0.030	0.067	1.774	0.137	—	—	—	—	—

(continued)

Table 2. (continued)

Item	Discrimination	Threshold/	SE	Diff. thresholds/	SE of diff.	E.S. convention	Z-test	Wald test	df	Equidistance index
Diff-Thr. Item 7		—	—	–0.968	0.207	Large	–4.676***	21.863***	1	0.968
Item 8	1.365	–2.747	0.270	—	—	—	—	—	—	—
Threshold 1–2		–1.343	0.138	1.404	0.194	—	—	—	—	—
Threshold 2–3		0.379	0.095	1.722	0.162	—	—	—	—	—
Diff-Thr. Item 8		—	—	–0.318	0.197	Small	—1.611	2.596	1	0.318

Note. The Wald tests run with 1 *df* in separate runs using Mplus 8.8. These results were supplemented with *Z*-tests and effect size conventions as described in the text. The equidistance index (Spratto, 2018) quantifies the difference of differences that is between pairs of adjacent thresholds (1, 2 vs. 2, 3). E.S. Conventions followed the guidelines of Dorans and Holland (1992) see also Lin and Lin (2013). *df* = degrees of freedom; Diff-Thr. = difference between adjacent thresholds; equidistance would be manifested with difference estimates equal to zero. PYDS = Positive Youth Development Scale.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

models that are appropriate for continuous data. The present study further provides the means to evaluate the interval scaling assumption that underlies continuous data using the popular software Mplus. Extension of the present study could include the creation of a modified GRM in which response options are equidistant but such an extension is currently not available in commercial statistical packages. The authors advise caution, however, when using the GRM model with small samples. For example, Jiang et al. (2016) showed that a sample size of n = 500 would suffice for most conditions they tested, producing good recovery of parameter estimates; an exception was very large models for which an n = 1,000 was recommended.

In the future, it will be important to evaluate factors that may alleviate the magnitude of threshold non-equivalence. For example, tangential item characteristics (Tourangeau et al., 2000), vague wording (Bass et al., 1974; Spratto, 2018), negatively worded items (Barnette, 2000; Coleman, 2013; Corwyn, 2000), emotionality elicited items (Tourangeau & Smith, 1996), the presence of a neutral option (Nowlis et al., 2002), clear labeling (Borgers et al., 2003), social presence (Tourangeau et al., 2003), and participant motivation (Krosnick, 1991), all contribute to some degree in altering participants' response behaviors.

Appendix A. Mplus 8.8 Syntax for Fitting the GRM Model to the Caring Subscale of the PYDS.

Mplus Syntax	Explanation
TITLE: Fitting the GRM model and interval scaling	! Title of model
DATA: FILE IS ncare1.dat;	! Data file in ASCII format
VARIABLE: NAMES ARE y1-y8;	! Variable names
usevariables are y1-y8;	! Variables to be used in the model
categorical are y1-y8;	! Variables defined as categorical
analysis: ESTIMATOR = ML; Link is logit;	! Type of estimator
MODEL:	! Model command
f1 BY y1-y8* (Li1-Li8);	! Estimating slopes of one factor model
[y1$1-y8$1] (T1_i1-T1_i8);	! 1st threshold estimation of items 1–8
[y1$2-y8$2] (T2_i1-T2_i8);	! 2nd threshold estimation of items 1–8
[y1$3-y8$3] (T3_i1-T3_i8);	! 3d threshold estimation of items 1–8
f1* (Fvar);	! Variance of factor estimated at first
[f1*] (Fmean);	! Factor mean estimated at first
OUTPUT: Residual stdyx tech10;	! Additional evaluative information
Plot: Type is plot1; Type is plot2; Type is plot3;	! Requesting IRT plots
MODEL CONSTRAINT:	! Model constraint command
Fvar=1;	! Factor variance constrained to 1
Fmean=0;	! Factor mean constrained to 0
NEW(A_i1-A_i8 B1_i1-B1_i8 B2_i1- B2_i8 B3_i1-B3_i8	! Labels of newly created parameters
D12_i1-D12_i8 D34_i1-D34_i8	! Labels of threshold estimates
Dt1-Dt8);	! Labels of difference threshold estimates

(continued)

Appendix A. (continued)

Mplus Syntax	Explanation
DO(1, 8) A_i#=Li#*sqrt(Fvar);	! Estimating item discriminations
DO(1, 8) B1_i#=(T1_i#-(Li#*Fmean))/ (Li#*sqrt(Fvar));	! Estimating 1st threshold levels in logit
DO(1, 8) B2_i#=(T2_i#-(Li#*Fmean))/ (Li#*sqrt(Fvar));	! Estimating 2nd threshold levels in logit
DO(1, 8) B3_i#=(T3_i#-(Li#*Fmean))/ (Li#*sqrt(Fvar));	! Estimating 3d threshold levels in logit
! Item 1	! For Item 1
D12_i1=B2_i1-B1_i1;	! Difference between thresholds 1 and 2
D34_i1=B3_i1-B2_i1;	! Difference between thresholds 2 and 3
.	
.	
! Item 8	! For Item 8
D12_i8=B2_i8-B1_i8;	! Difference between thresholds 1 and 2
D34_i8=B3_i8-B2_i8;	! Difference between thresholds 2 and 3
Model Test:	! Wald test for evaluating equivalence
0=D12_i1-D34_i1;	! between adjacent thresholds, i.e., 1 ! and 2 vs. 2 and 3 for the 1st item. ! Wald tests for each item require ! additional runs in Mplus.
	! Differences in thresholds using Z-tests
dt1=D12_i1-D34_i1;	! Difference threshold estimates in Item 1
dt2=D12_i2-D34_i2;	! Difference threshold estimates in Item 2
dt3=D12_i3-D34_i3;	! Difference threshold estimates in Item 3
dt4=D12_i4-D34_i4;	! Difference threshold estimates in Item 4
dt5=D12_i5-D34_i5;	! Difference threshold estimates in Item 5
dt6=D12_i6-D34_i6;	! Difference threshold estimates in Item 6
dt7=D12_i7-D34_i7;	! Difference threshold estimates in Item 7
dt8=D12_i8-D34_i8;	! Difference threshold estimates in Item 8

Note. More model test command runs need to follow for testing the equivalence in each item as one Wald test can be conducted at a time. Commands in the "Model Constraint" section are used to evaluate the interval scaling assumption. ASCII = American Standard Code for Information Interchange; GRM = graded response model; PYDS = Positive Youth Development Scale; ML = maximum likelihood; IRT = item response theory.

Appendix B. Original Caring Items From the Positive Youth Development Scale.

Please rate how strongly you agree or disagree with the following statements.	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1. When there is a need I offer assistance whenever I can	☐	☐	☐	☐	☐
2. It is easy for me to consider the feelings of others*	☐	☐	☐	☐	☐
3. I care about how my decisions affect other people*	☐	☐	☐	☐	☐

(continued)

Appendix B. (continued)

Please rate how strongly you agree or disagree with the following statements.	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
4. I try to encourage others when they are not as good at something as me*	☐	☐	☐	☐	☐
5. Other people's feelings matter to me*	☐	☐	☐	☐	☐
6. I can be counted on to help if someone needs me.	☐	☐	☐	☐	☐
7. I care about the feelings of my friends*	☐	☐	☐	☐	☐
8. When one of my friends is hurting, I hurt too.	☐	☐	☐	☐	☐

Note. Note that the mid option (neither/nor) was not utilized in the Arabic version of the instrument with the number of options being only 4. *=code to signal freely estimated parameter.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The present study was funded by ETEC, Saudi Arabia.

ORCID iD

Georgios Sideridis https://orcid.org/0000-0002-4393-5995

References

Akdogan, R., & Turkum, A. S. (2018). Insight scale for nonclinical university students: Validity and reliability analysis. *Measurement and Evaluation in Counseling and Development*, *51*, 250–262.

Aldawsari, H., Laux, J., Dari, T., & Gaballa, H. (2021). The revised Arabic Schwartz outcome scale-10 (SOS-10-AR). *Measurement and Evaluation in Counseling and Development*, *54*, 120–129.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, *38*, 665–680. https://doi.org/10.1177/001316447803800308

Arnold, M. E., Nott, B. D., & Meinhold, J. L. (2012). The Positive Youth Development Inventory Full Version. © Oregon State University.

Autin, K., Blustein, D., Duffy, R., Gensmer, N., Douglass, R., & Allan, B. (2019). The development and initial validation of need satisfaction scales within the psychology of working theory. *Journal of Counseling Psychology*, *66*, 195–209.

Babakus, E., Ferguson, C. E. J., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*, 222–228.

Bardoshi, G., Erford, B., & Jang, H. (2019). Psychometric synthesis of the counselor burnout inventory. *Journal of Counseling & Development*, *97*, 195–208.

Barnette, J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*, 361–370.

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology*, *59*(3), 313–320.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Blau, G., & Dimino, J. (2019). Prepared for counseling: Introducing a short scale and correlates. *Measurement and Evaluation in Counseling and Development*, *52*, 274–283.

Bloom, Z., & Dillman-Taylor, D. (2020). The online dating intensity scale: Exploratory factor analysis in a sample of emerging adults. *Measurement and Evaluation in Counseling and Development*, *53*, 1–16.

Bollen, K. A., & Barb, K. H. (1981). Pearson's R and coarsely categorized measures. *American Sociological Review*, *46*, 232–239.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Borgers, N., Hox, J., & Sikkel, D. (2003). Response quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research*, *15*(1), 83–94.

Brislin, W. R. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185–216.

Cho, Y., Choi, Y., Kim, S., & Hong, S. (2018). Factor structure and other psychometric properties of the social phobia inventory in Korean samples. *Measurement and Evaluation in Counseling and Development*, *51*, 263–280.

Cimsir, E., & Akdogan, R. (2021). Childhood emotional incest scale (CEIS): Development, cross-validation and reliability. *Journal of Counseling Psychology, 68*, 98–111

Coleman, C. (2013). *Effects of negative keying and wording in attitude measures: A mixed-methods study* [Unpublished doctoral dissertation, James Madison University].

Corwyn, R. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality*, *34*, 357–379.

Dillman, D., Bratton, S., & Henson, R. (2019). Confirming the constructs of Adlerian personality priority assessment. *Measurement and Evaluation in Counseling and Development*, *52*, 191–206.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*, 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Dorans, N. J., & Holland, P. W. (1992, October). *DIF detection and description: Mantel-Haenszel and standardization* [Paper presentation]. *Presented at the educational testing service/AFHRL conference*, Princeton, NJ, United States.

Erford, B., Sriken, J., Sherman, M., Hibbs, J., Smith, H., Kipper-Smith, A., & Niarhos, F. (2021). Psychometric analysis, internal structure, and measurement invariance of the alcohol use disorders identification test (AUDIT) scores from a large university sample. *Measurement and Evaluation in Counseling and Development*, *54*, 188–205.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625–641.

Fu, M., & Zhang, L. (2019). Developing and validating the career personality styles inventory. *Measurement and Evaluation in Counseling and Development*, *52*, 38–51.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*(3), 564–567.

Ganho-Avila, A., Moura-Ramos, M., Goncalves, O., & Almeida, J. (2019). Measuring vulnerability to anxiety: Factorial structure, reliability, validity, and discriminatory accuracy of the anxiety sensitivity index-3-pt. *Measurement and Evaluation in Counseling and Development*, *52*, 223–238.

Garifio, J., & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *33*(3), 106–116.

Ghabrial, M., & Andersen, J. (2021). Development and initial validation of the queer people of color identity affirmation scale. *Journal of Counseling Psychology*, *68*, 38–53.

Ghosh, A., Niileksela, C., Parham, A., & Janis, R. (2021). Investigating factorial invariance of the counseling center assessment of psychological symptoms-34 (CCAPS-34) with military and nonmilitary students. *Measurement and Evaluation in Counseling and Development*, *54*, 42–55.

Gonzalez, E., Sperandio, K., Mullen, P., & Tuazon, V. (2021). Development and initial testing of the multidimensional cultural humility scale. *Measurement and Evaluation in Counseling and Development*, *54*, 56–70.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling*, *4*, 108–120.

Greene, J. (2019). The multicultural school counseling behavior scale: Development, psychometrics and use. *Professional School Counseling, 22(1)*, 1–10.

Griffin, B., Worthington, E., Hook, J., Davis, D., & Maguen, S. (2018). Development of the self-forgiveness dual-process scale. *Journal of Counseling Psychology*, *65*, 715–726.

Halamova, J., Kanovsky, M., Petrocchi, N., Moreira, H., Lopez, A., Bamett, M., Yang, E., Benda, J., Brahler, E., Zeng, X., & Zenger, M. (2021). Factor structure of the self-compassion scale in 11 international samples. *Measurement and Evaluation in Counseling and Development*, *54*, 1–23.

Henkel, R. E. (1975). Part-whole correlations and the treatment of ordinal and quasi interval data as interval data. *The Pacific Sociological Review*, *18*(1), 3–26. https://doi.org/10.2307/1388217

Hensler, C., & Stipak, B. (1979). Estimating interval scale values for survey item response categories. *American Journal of Political Science*, *23*(3), 627–649. https://doi.org/10.2307/2111033

Hiles, A., Dandy, K., Martino, M., Howard, J., & Cross, D. (2019). An evaluation of the psychometric properties of the attachment behavior checklist. *Measurement and Evaluation in Counseling and Development*, *52*, 171–190.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 344–364. https://doi.org/10.1080/10705519809540111

Jakobsson, U. (2004). Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences*, *18*, 437–440.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*, 1212–1218.

Jiang, S., Wang, C., & Weiss, D. J. (2016) Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model. *Frontiers in Psychology*, *7*, 109. doi: 10.3389/fpsyg.2016.00109

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, *48*, 398–407.

Johnson, D., & Karcher, M. (2019). Validity evidence for a state-specific assessment of in-session counselor empathy: The state-empathic concern scale. *Measurement and Evaluation in Counseling and Development*, *52*, 284–296.

Johnson, D., Knight, D., & McHugh, K. (2021). Score reliability and validity evidence for the state-interpersonal reactivity index: A multidimensional assessment of in-session counselor empathy. *Measurement and Evaluation in Counseling and Development*, *54*, 24–41.

Kim, Y., Wang, Y., & Li, X. (2021). Examination of the identify style inventory with racial/ethnic minority college students: Revision of factor structure and psychometrics. *Measurement and Evaluation in Counseling and Development*, *54*, 173–187.

Kivlighan, D. M., Adams, M., Tao, K., Drinane, J., & Owen, J. (2019). Construction and validation of the multicultural orientation inventory-group version. *Journal of Counseling Psychology*, *66*, 45–55.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research*, *39*(2), 121–123. https://doi.org/10.1097/00006199-199003000-00019

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Lau, J., Ng, K., & Vallett, D. (2019). The counseling training environment scale: Initial development and validity of a self-report measure to assess the counseling training environment. *Measurement and Evaluation in Counseling and Development*, *52*, 255–273.

Lee, D., Gaskin-Wasson, A., Jones, C., Harrell, S., Banks, K., Kohn-Wood, L., Sellers, R., & Neblett, E. (2019). The daily life experiences scale: Factor structure, reliability, validity, and measurement invariance for African American males and females. *Measurement and Evaluation in Counseling and Development*, *54*, 206–218.

Lee, J., & Shin, Y. (2019). Experience in close relationships scale-short version (ECR-S) validation with Korean college students. *Measurement and Evaluation in Counseling and Development*, *52*, 119–127.

Levant, R., & Parent, M. (2019). The development and evaluation of a brief short form of the normative male alexithymia scale (NMAS-BF). *Journal of Counseling Psychology*, *66*, 224–233.

Levant, R., Parent, M., Mahalik, J., McDermott, R., Alshabani, N., & Hammer, J. (2020). Development and evaluation of a new short form of the conformity to masculine norms inventory (CMNI-30). *Journal of Counseling Psychology*, *67*, 622–636.

Likert, R. (1932). The method of constructing an attitude scale. In S. R. Houston, J. Schmid, R. Lynch, & W. Duff, Jr. (Eds.), *Methods and techniques in business research* (pp. 53–65). MSS Information Corporation.

Lim, E., & Kim, S. (2020). A validation of a multicultural competency measure among South Korean counselors. *Journal of Multicultural Counseling & Development*, *48*, 15–29.

Liu, T., Maffini, C., Iwamoto, D., Wong, Y. J., & Mitts, N. (2018). Gendered racism scales for Asian American men: Scale development and psychometric properties. *Journal of Counseling Psychology*, *65*, 556–570.

Lin, P. Y., & Lin, Y. C. (2013). Examining student factors in sources of setting accommodation DIF. *Educational and Psychological Measurement*, *74*(5), 759–794.

Lu, J., Woo, H., & Huffman, K. (2018). Spiritual competency scale: A confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, *51*, 219–234.

Ludlow, L., Matz-Costa, C., & Klein, K. (2019). Enhancement and validation of the productive engagement portfolio-scenario (PEP-S8) scales. *Measurement and Evaluation in Counseling and Development*, *52*, 15–37.

Luo, Y., Watson, J., & Lenz, S. (2021). Development and initial validation of a social media identity distress scale among emerging adults. *Measurement and Evaluation in Counseling and Development*, *54*, 141–155.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Martens, M. P. (2005). The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist*, *33*(3), 269–298.

Martin, J., Zamboanga, B., Haase, R., & Buckner, L. (2020). Measurement invariance of the protective behavioral strategies scale across racial groups. *Measurement and Evaluation in Counseling and Development*, *53*, 17–33.

Mazahreh, L., Stoltz, K., & Wolff, L. (2019). Validation of the career adapt-abilities scale in the Hashemite kingdom of Jordan. *Measurement and Evaluation in Counseling and Development*, *52*, 108–118.

Moate, R., Gnilka, P., West, E., & Rice, K. (2019). Doctoral student perfectionism and emotional well-being. *Measurement and Evaluation in Counseling and Development*, *52*, 145–155.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muthen, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–189.

Muthén, L. K., & Muthén, B. O. (1998–2018). *Mplus user's guide*.

Nowlis, S. M., Kahn, B. E., & Dhar, R. (2002). Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments. *Journal of Consumer Research*, *29*, 319–334.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

O'Brien, R. M. (1979). The use of Pearson's r with ordinal data. *American Sociological Review*, *44*(5), 851–857. https://doi.org/10.2307/2094532

Oh, S., & Butler-Shillingford, A. (2019). The client assessment of multicultural competent behavior (CAMCB): Development and validation. *Measurement and Evaluation in Counseling and Development*, *54*, 71–89.

Pederson, C., Gorman-Ezell, K., Mayer, G., & Brookings, J. (2021). Development and preliminary validation of a tool for screening suicide risk in chronically III women. *Measurement and Evaluation in Counseling and Development*, *54*, 130–140.

Perez-Rojas, A., Lockard, A., Bartholomew, T., & Gonzalez, J. (2019). Development and initial validation of the therapist cultural comfort scale. *Journal of Counseling Psychology*, *66*, 534–549.

Pohl, N. F. (1981). Scale considerations in using vague quantifiers. *The Journal of Experimental Education*, *49*(4), 235–240. https://doi.org/10.1080/00220973.1981.11011790

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, *46*, 273–286.

Poynton, T., Ruiz, B., & Lapan, R. (2019). Development and validation of the college admissions knowledge evaluation. *Professional School Counseling, 22(1b)*, 1–6.

Pozza, A., Barcaccia, B., & Dettoer, D. (2019). Psychometric evaluation of the Italian obsessive compulsive intentory-child version: Factor structure and predictive validity at one-year follow-up in adolescents. *Measurement and Evaluation in Counseling and Development*, *52*, 239–254.

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*, 354–373. https://doi.org/10.1037/a0029315

Rigdon, E. E., & Fergusson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, *28*, 491–497.

Rowan-Kenyon, H., McCready, A., Aleman, A., & Barone, N. (2021). Measuring racist aggressions on social media and the effects on U.S. college students of color: An instrument validation. *Measurement and Evaluation in Counseling and Development*, *54*, 156–172.

Samejima, F. (1969). *Estimation of latent ability a response pattern of graded scores* (Psychometrika Monograph Supplement, No. 17). https://doi.org/10.1007/BF03372160

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for statistics in covariance structure analysis (UCLA Statistics Series #2). Los Angeles, CA: University of California.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications to developmental research* (pp. 399–419). SAGE.

Shea, M., Nguyen, K., Wong, Y. J., & Gonzalez, P. (2019). College students' barriers to seeking mental health counseling: Scale development and psychometric evaluation. *Journal of Counseling Psychology*, *66*, 626–639.

Shin, R., Lu, Y., Welch, J., Sharma, R., Vernay, C., Yee, S., & Smith, L. (2018). The development and validation of the contemporary consciousness measure II. *Journal of Counseling Psychology*, *65*, 539–555.

Simons, J. (2018). School counselor sexual minority advocacy competence scale (SCSMACS): Development, validity, and reliability. *Professional School Counseling, 21(1)*, 1–14.

Spector, P. E. (1980). Ratings of equal ad unequal response choice intervals. *The Journal of Social Psychology*, *112*(1), 115–119. https://doi.org/10.1080/00224545.1980.9924303

Spratto, E. M. (2018). *In search of equality: Developing an equal interval Likert response scale* [Doctoral dissertations]. https://commons.lib.jmu.edu/diss201019/172

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

Sullivan, G., & Artino, A. R. Jr. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542.

Swank, J., Limberg, D., & Liu, R. (2020). Development of the altruism scale for children: An assessment of caring behaviors among children. *Measurement and Evaluation in Counseling and Development*, *53*, 34–43.

Tadlock-Marlo, R., & Hill, N. (2019). Validation and psychometric properties of one school, many differences. *Measurement and Evaluation in Counseling and Development*, *52*, 156–170.

TaeHyuk-Keum, B., Brady, J., Sharma, R., Lu, Y., Kim, Y., & Thai, C. (2018). Gendered racial microaggressions scale for Asian American women: Development and initial validation. *Journal of Counseling Psychology*, *65*, 571–585.

Toland, M., Li, C., Kodet, J., & Reese, R. (2021). Psychometric properties of the outcome rating scale: An item response theory analysis. *Measurement and Evaluation in Counseling and Development*, *54*, 90–105.

Tourangeau, R., Couper, M. P., & Steiger, D. M. (2003). Humanizing self-administered surveys: Experiments on social presence in Web and IVR surveys. *Computers in Human Behavior*, *19*, 1–24.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275–304.

Trub, L., & Barbot, B. (2020). Great escape or path to self-expression?: Development and validation of a scale of motivations for text messaging. *Measurement and Evaluation in Counseling and Development*, *53*, 44–61.

Tsaousis, I., Sideridis, G. D., & Alahmadi, M. T. S. (2021). Applying a multidimensional item response theory approach in validating the dimensionality of the positive youth development scale. *Journal of Psychoeducational Assessment*, *397*(7), 861–873. https://doi.org/10.1177/07342829211002332

Veronese, G., & Pepe, A. (2019). Using the posttraumatic growth inventory-short form with Palestinian helpers living in conflict areas. *Measurement and Evaluation in Counseling and Development*, *52*, 207–221.

Vigderhous, G. (1977). The level of measurement and ''permissible'' statistical analysis in social research. *The Pacific Sociological Review*, *20*(1), 61–72. https://doi.org/10.2307/1388904

Waldrop, D., Reschly, A., Fraysier, K., & Appleton, J. (2019). Measuring the engagement of college students: Administration format, structure, and validity of the student engagement instrument-college. *Measurement and Evaluation in Counseling and Development*, *52*, 90–107.
Wang, L., Koay, E., Wei, M., Lo, M., & Lee, M. (2019). The development and validation of the emotional cultivation scale: An east Asian cultural perspective. *Journal of Counseling Psychology*, *66*, 409–423.
Watson, L., Allen, L., Flores, M., Serpe, C., & Farrell, M. (2019). The development and psychometric evaluation of the trans discrimination scale: TDS-21. *Journal of Counseling Psychology*, *66*, 14–29.
Watson, J., Prosek, E., & Giordano, A. (2020). Investigating psychometric properties of social media addiction measures among adolescents. *Journal of Counseling & Development*, *98*, 458–466.
Xavier, A., Cunha, M., & Pinto-Gouveia, J. (2019). Validation of the risk-taking and self-harm inventory for adolescents in a Portuguese community sample. *Measurement and Evaluation in Counseling and Development*, *52*, 1–14.
Young, A., & Bryan, J. (2018). The school counselor leadership survey: Confirmatory factor analysis and validation. *Measurement and Evaluation in Counseling and Development*, *51*, 235–249.
Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289–309.
Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modeling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology*, *64*, 107–133.
Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, *34*, 390–400.
Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.