

Chapter 2

Thurstonian Item Response Theory and an Application to Attitude Items

Edward H. Ip

Abstract The assessment of attitudes has a long history dating back at least to the work of Thurstone. The Thurstonian approach had its “golden days,” but today it is seldom used, partly because judges are needed to assess the location of an item, but also because of the emergence of contemporary tools such as the IRT. The current work is motivated by a study that assesses medical students’ attitudes toward obese patients. During the item-development phase, the study team discovered that there were items on which the team members could not agree with regard to whether they represented positive or negative attitudes. Subsequently, a panel of $n = 201$ judges from the medical profession were recruited to rate the items, and the responses to the items were collected from a sample of $n = 103$ medical students. In the current work, a new methodology is proposed to extend the IRT for scoring student responses, and an affine transformation maps the judges’ scale onto the IRT scale. The model also takes into account measurement errors in the judges’ ratings. It is demonstrated that the linear logistic test model can be used to implement the proposed Thurstonian IRT approach.

Keywords Item response theory • Likert scaling • Linear logistic test model • Attitudes toward obese persons • Equal-appearing interval scaling

2.1 Introduction

Together with the Guttman scale, Thurstone and Likert scaling are perhaps the most prominently featured and researched scaling techniques in the history of psychological measurement, especially in the assessment of attitudes. Historically, Thurstone was one of the first quantitative psychologists to set his sights on the development of a theory for psychological scaling (Thurstone 1925, 1928). Thurstone’s pioneer work on conception of attitude was based on the assessment of subjective attitudinal responses. The covert responses—or a sample of them—are

E.H. Ip (✉)

Department of Biostatistical Sciences, Wake Forest School of Medicine,
Medical Center Blvd., WC23, Winston-Salem, NC 27157, USA

e-mail: eip@wakehealth.edu

linguistically represented in the form of opinion statements, which can then be located on a single evaluative dimension (Ostram 1989). Based on the principle of comparative judgment, Thurstone developed several scaling methods, of which the best known is the equal-appearing interval scale (Thurstone and Chave 1929). Given a collection of items, each of which contains a statement concerning the psychological construct of interest, the technique consists of two steps.

First, a panel of judges is recruited to rate the items in terms of their favorability to the construct of interest. Thurstone suggested using integral values of 1–11 for the rating scale. The 11-point scale then becomes the psychological continuum on which the statements have been judged, and the distribution of judgments obtained is used to calculate a typical value, which can then be taken as the scale-value of the statement on the 11-point psychological continuum. The value could be the median or the mean of the judgment distribution, and descriptive statistics such as standard deviations and the interquartile range are then used to eliminate questions that have overly dispersed judgment scores. Ideally, the equal-appearing interval scale is established by a final collection of items with small dispersions so that the scale-value of the statements on the psychological continuum are relatively equally spaced. In the second step, the statements are presented to subjects with instructions to indicate those with which they are willing to agree and those with which they disagree. The attitude score for a subject is based on the mean or the median of the scale-values of the statements agreed with. In other words, if the responses are dichotomously coded as 1 for Agree and 0 for Disagree, then the attitude score is an average of a weighted combination of the response categories, of which the weights are the scale-score.

One of the most fascinating aspects of Thurstone's scaling procedure is that the scale is determined by expert judges on a unidimensional continuum and that the operating characteristic of a Thurstone item may reflect either an underlying dominant-response process or an ideal-point process (Coombs 1964; Roberts and Laughlin 1996). In the most common form of the dominance mechanism, respondents and items are represented by positions on a latent trait, and the responses are determined by a comparison process: if the respondent's trait value is greater than the item-trait value, then the response to the item is positive; otherwise, the response is negative. The item-characteristic curve (ICC) of the item response for a dominant-response process is monotone and can be well captured by existing item response theory (IRT; Lord 1980) models. An example of a monotone ICC for equal-appearing interval scaling is the Sickness-Impact Profile (SIP; Bergner et al. 1981). Judges rated the SIP items on the severity of the dysfunction described in an item on an equal-interval 11-point scale. The end points were labeled "minimally dysfunctional" and "severely dysfunctional" to provide meaningful referents. An item concerning how sickness impacts work is: "I act irritable and impatient with myself—for example, talk badly about myself, swear at myself, blame myself for things that happen." A monotone ICC for this item implies that a respondent with

a higher SIP trait value (more dysfunctional) is more likely to endorse this item than someone with a lower SIP trait value (less dysfunctional). For an empirical comparison between IRT scaling and Thurstone scaling in education, see Williams et al. (1998).

The Thurstone scaling procedure could also be used to describe an ideal point-response process, a model commonly used in attitudinal measurement of political and social views. Like the dominant-response process, the ideal-point process postulates that the individual's response also depends on the relative position of the person's trait value and the position of the item on the scale. However, a respondent in an ideal-point process is more likely to endorse statements that have trait values close to the respondent's. Thus, the ICC from an ideal-point process is not monotonic with respect to the trait and typically has a single peak at the location of the item. These models are often referred to as unfolding models in the IRT literature. An example of an unfolding item is a well-known General Social Survey (GSS) item on legalized abortion. The respondent in the GSS is asked when legalized abortion is allowed on a collection of seven conditions such as: "The family has a very low income and cannot afford any more children" and "The woman wants it for any reason." For respondents who hold a more centralist view about legalized abortion, the likelihood of endorsing the former statement would be higher than it would be for those who hold a liberal view about abortion as well as those who are strong anti-abortion.

In this paper, we only focus on Thurstone's equal-appearing scaling methods for items that do not fold—or items that are supposed to follow a dominant-response process so that their ICCs are monotonic. We argue that the equal-appearing scaling method is a way to set scales according to experts' views of the construct of interest and that it could be operationalized through IRT models in which the location parameter of an item can be obtained by careful scaling of the judges' ratings. The extent to which the judges disagree on the location of an item can be incorporated into the IRT model by assuming that the rating scores from the sample of judges are normally distributed with a mean m and a standard deviation σ , both of which could be directly estimated from the judges' data. As such, the proposed model can be viewed as an IRT implementation for equal-appearing scaling, which is distinct from the Thurstonian item response model proposed in Brown and Maydeu-Olivares (2012). We further demonstrate that the uncertainty associated with the judges' ratings would lead to an attenuation of the slope of the ICC, which, in modern IRT language, means that the information contained in the item is less than 1 at the same scale location but has a steeper slope. Also, we show that through a convolution technique the proposed Thurstonian IRT model can be solved using the estimation procedure for the linear logistic test model (LLTM; Fischer 1973).

The remainder of this paper is organized as follows: first, we describe the Thurston IRT model, then we illustrate the proposed model using a data set collected from a study of attitudes. Finally, we conclude with a discussion.

2.2 Thurstonian IRT: Method

We begin with a simple Rasch model:

$$P(Y_{ij} = 1 | \theta_j) = \frac{\exp(\theta_j + b_i)}{1 + \exp(\theta_j + b_i)}, \quad (1)$$

where Y_{ij} is the binary response of individual j to item i , with 1 indicating a correct or positive response; θ_j is the latent trait for individual j ; and b_i is the intercept parameter for item i or the individual. We further assume that the intercept parameter b_i is a function of item attributes \underline{w}_i and the judge's rating, which has a mean m_i and variance σ_i^2 . Specifically, we write:

$$b_i = \eta_1^T \underline{w}_i + \eta_2(m_i + \varepsilon_i), \quad \varepsilon_i \sim N(0, \sigma_i^2), \quad (2)$$

where η denotes regression coefficients.

$$P(Y_{ij} = 1 | \theta_j, \varepsilon_i) = \frac{\exp[\theta_j + \eta_1^T \underline{w}_i + \eta_2(m_i + \varepsilon_i)]}{1 + \exp[\theta_j + \eta_1^T \underline{w}_i + \eta_2(m_i + \varepsilon_i)]}, \quad (3)$$

$$\theta_j \sim N(0, 1), \quad \varepsilon_i \sim N(0, \sigma_i^2).$$

In other words, we have

$$P(Y_{ij} = 1 | \theta_j, \varepsilon_i) = \frac{\exp[\theta_j + b'_i + \eta_2 \varepsilon_i]}{1 + \exp[\theta_j + b'_i + \eta_2 \varepsilon_i]}, \quad (4)$$

where $b'_i = \eta_1^T \underline{w}_i + \eta_2 m_i$.

By integrating out the error term $\eta_2 \varepsilon_i$ through a convolution technique (Zeger et al. 1988; Caffo et al. 2007; Ip 2010), we now have

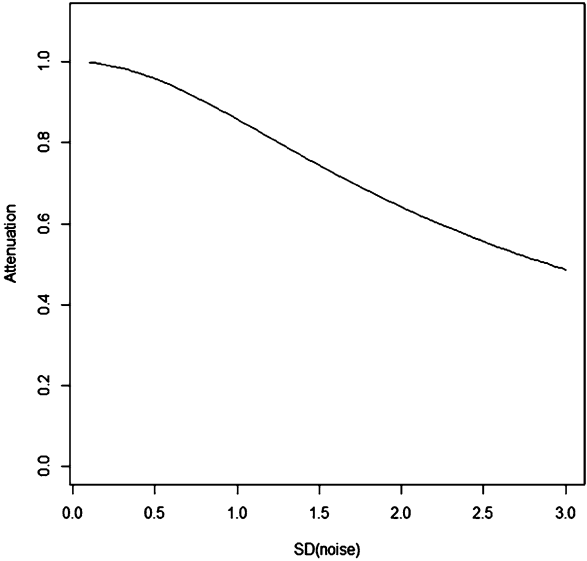
$$P(Y_{ij} = 1 | \theta_j) = \frac{\exp[a_i^* \theta_j + b_i^*]}{1 + \exp[a_i^* \theta_j + b_i^*]}, \quad (5)$$

where $a_i^* = \lambda_{\text{logit}}(a_{i1} + \frac{\eta_2 \rho \sigma_i}{\sigma_1})$, $b_i^* = \lambda_{\text{logit}} b'_i$, $\lambda_{\text{logit}} = [k^2 \eta_2^2 (1 - \rho^2) \sigma_i^2 + 1]^{-1/2}$, and $k = 16\sqrt{3}/(15\pi) = 0.588$. The factor a_i^* represents an attenuation factor for the slope of θ , which is assumed to be 1.0 in a Rasch model, and ρ represents the correlation between ε and θ , which is set to zero.

Figure 2.1 shows the change in attenuation as a function of the standard deviation of the measurement error. Generally speaking, when the noise level (measurement error) increases, the attenuation factor becomes smaller and varies almost linearly

from no attenuation ($=1.0$) to a value of 0.5. Notably, the graphs show that attenuation is approximately 0.8 when the noise level ($SD = 1$) reaches the level of the signal ($SD = 1$). We call the model specified by Eqs. (4) and (5) the Thurstonian LLTM model.

Fig. 2.1 Attenuation factor as a function of the standard deviation of the judges' ratings



2.3 Real Data Example

2.3.1 Data

The data were a subset of data collected from a recent study on the development of a curriculum for medical school students for counseling obese patients. The Nutrition, Exercise, and Weight Management (NEW) study collected attitude data using an instrument—the NEW Attitude Scale (Ip et al. 2013)—which comprises 31 items measuring attitudes across three domains: nutrition, exercise, and weight management. Examples of items include “I do feel a bit disgusted when treating a patient who is obese” (Item 23), and “The person and not the weight is the focus of weight-management counseling” (Item 25). In the item-development process, the study team had a consensus view for some items but divergent views for others. An example of a consensus item was “Overweight individuals tend to be lazy about exercise” (Item 13), which the team agreed represented an unfavorable

attitude. An item that solicited divergent views was “Patients are likely to follow an agreed-upon plan to increase their exercise” (Item 10). Some tended to feel that an endorsement of the item suggested a favorable attitude because the physician sounded positive about the outcome, but others argued that the item should be viewed negatively because the physician might not appreciate the challenges that an obese person encountered when prescribed an exercise program. The study team decided to use the Thurstonian approach of soliciting judges’ opinions about the positivity/negativity of the items. A total of 201 judges (approximately 50% clinically focused and the remaining research focused) rated the items. A sample of N = 103 medical students completed the instrument. Using the scores that were derived from traditional Thurstone scaling, the test–retest reliability of the instrument was 0.89 (N = 24). Pearson correlations between two other anti-obesity measures were the Anti-Fat Attitudes Questionnaire (AFA; Lewis et al. 1997) and the Beliefs About Obese Persons Scale (BAOP; Allison et al. 1991) were -0.47 and 0.23, respectively. This shows satisfactory convergent validity with existing measures of attitudes toward obese individuals. A full report about the validation of the instrument can be found in Ip et al. (2013).

To illustrate the range of concordance in judges’ ratings across items, we used two items as examples. Figures 2.2 and 2.3 show, respectively, the distributions of ratings for Item 23 and Item 25. The former item has a relatively high level of consensus as being indicative of an unfavorable attitude, as demonstrated by the small standard deviation (SD = 0.8). In contrast, Item 25 exhibits high variance in the judges’ ratings (SD = 2.2).

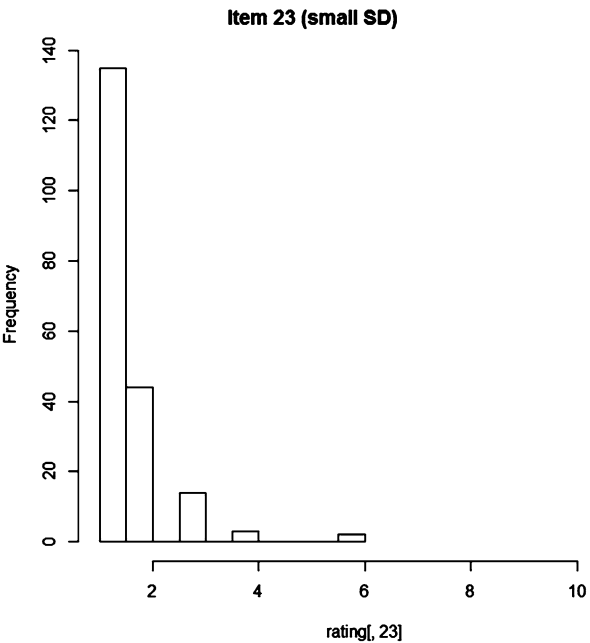
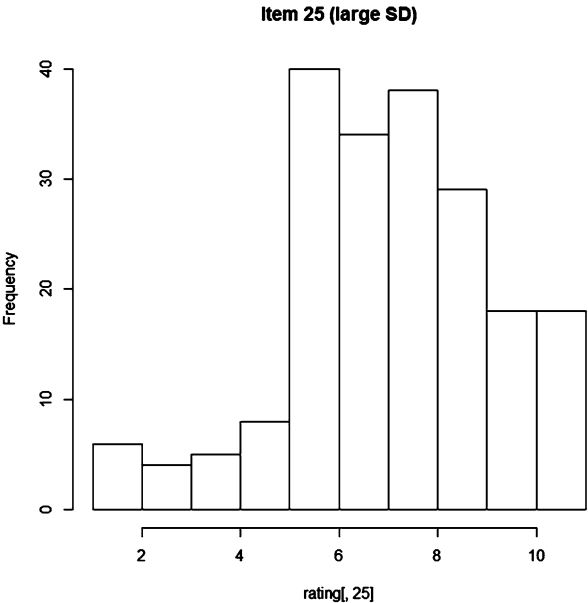


Fig. 2.2 Distribution of judges’ ratings for Item 23

Fig. 2.3 Distribution of judges' ratings for Item 25



Besides the three domains (nutrition, exercise, and weight management) that defined the items, it was expected that some items also carried common characteristics. For example, there were items across the three domains that were related to counseling, and there were also items that were related to motivation of the patient in dieting, exercise, and weight loss. Therefore, we also conducted a factor analysis to extract factors that explained a large proportion of the variance of the items.

We used the Thurstonian LLTM described above to estimate the model parameters, and in addition to the judges' ratings the following two covariates were included: the factor score of the item from factor analysis and the domain to which each item belonged. A standard LLTM program eRm (Mair and Hatzinger 2007) was used to estimate the parameters.

2.3.2 Results

The factor analysis resulted in three factors that can be interpreted as (1) a factor for counseling, (2) a factor for motivation of the patient, and (3) a factor for perception about external factors. Table 2.1 summarizes the results from the Thurstonian LLTM and reports the estimates and standard errors (SE). Except for the exercise domain (as compared with weight management), all of the predictors that were entered into the LLTM are significant. Specifically, judges' ratings tend to be highly significant, and each point increase in a judge's rating results in a change of -1.4 in the intercept parameter. Figure 2.4 shows the ICCs for two exemplifying items: the solid line

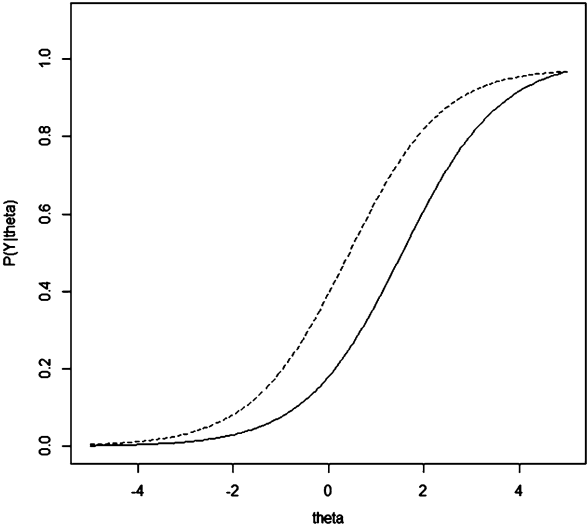
shows that of Item 23 (“Patients tend to be lazy about exercise”) and Item 14 (“Patients understand the connection between nutrition and cancer”). The ICC for Item 23 suggests that medical students with higher values on the NEW Attitude Scale are less likely to endorse this item than they are to endorse Item 14. Finally, the effect of measurement error on the attenuation within the LLTM appeared to be small. The attenuation factor for the items in the sample ranged from 0.96 to 0.99.

Table 2.1 Estimates and standard error for Thurstonian LLTM for NEW attitude data

Predictor	Estimate for η	SE
Factor 1	−1.528*	0.115
Factor 2	−0.85*	0.125
Nutrition	0.449*	0.104
Exercise	−0.04	0.102
Judges’ ratings	−0.143*	0.02

* $p < 0.01$
Factor 3 is the reference factor

Fig. 2.4 Item-characteristic curves for Item 23 (solid) and Item 14 (dashed)



2.4 Discussion

There is often misunderstanding and confusion in the literature about the use of the Likert scaling method (Likert 1932; Edwards and Kenney 1946). Partly because of the convenience of constructing items and scoring respondents, it is not uncommon to see confusion about the fundamental scaling idea behind the Likert method. In particular, one misconception about the Likert scale that is relevant to this paper is that using the Likert scale does not require a specific scaling procedure—i.e.,

calibrating the continuum of metric by identifying the locations of the items on the continuum because no judges are required. This is not true. Likert actually suggested more than one way of assigning scale values, and indeed there are at least three groups of persons that are capable to locating items on a continuum: (1) a panel of expert judges, (2) the test developers, and (3) the respondents. Thurstone relied on the first category, and Likert developed methods in using the other two categories of persons. To understand his notion of scaling, we need to briefly describe Likert's assumptions underlying his procedure. Instead of following Thurstone's approach of creating positional statements, Likert used the level of agreement with specific statements to measure attitudes. The "level of agreement" could be codified as Strongly Agree to Strongly Disagree, or as judgmental statements about actions concerning a given situation. In his study about racial attitudes among college students, one of the questions was: "In a community in which the negroes [sic] outnumber the whites, under what circumstances is the lynching of a negro [sic] justifiable?" The possible responses were: "(a) Never. (b) In very exceptional cases where an especially brutal crime against a white person calls for swift punishment. (c) As punishment for any brutal crime against a white person. (d) As punishment for any gross offense [felony or extreme insolence] committed against a white person. (e) As punishment for any act of insolence against a white person." It is difficult not to notice the similarity of these response categories to statements in a Thurstone scaling procedure. The response categories, when expressed in this form could be more appropriately called sub-statements. Indeed, Likert scaling corresponds to a scheme under which the test developers provide the rating for the sub-statements (e.g., see Massof 2002).

The argument that Likert scaling corresponds to a predetermined scale is based on the observation that Likert's "theory" of scaling assumes that attitudes in a population follow a normal distribution and that all items can be positioned on the continuum by assigning them sigma units, or what we call z-scores now. Instead of using continuous values, Likert argued that one could partition the continuum into response categories, each of which signified a level of intensity on the continuum. A critical step that Likert took was to assign ranks (1–5) to these intensity categories.

From the perspective of the Thurstone scaling procedure, Likert scaling is equivalent to assigning transformed z-scores (1–5) as scale values to the sub-statements in an item. If each sub-statement in an attitude instrument is treated as a statement on Thurstone's equal-appearing interval, there would be five distributions at five equally separated positions. In other words, Likert's scaling corresponds to a form of equal-appearing interval scaling in which 5 points are used instead of 11. The test developer assigns the scale value to each item, and it is assumed that the assignment is without error. Alternatively, Likert alluded to the use of the participants as rating "judges"—i.e., the intensity of an item is determined by how frequent high scorers endorse the item (Babbie 2008, p. 188). Thus, although Likert scaling creates the ordinal format in order to avoid the need for external judges when developing scales, the scaling of the items still has to come from somewhere—for example, either from a test developer or from the participants. Some criticized the Thurstone scaling procedure because while it is valid for judges it may not

be valid for participants. Yet, this is the whole point of Thurstone—the judges, presumably practitioners and researchers in the field, set the scale for a construct that they have all judged to be measurable using the proposed items. One can even argue that this scaling method would be a more relevant measure for a construct because a construct, after all, is an artifact conceived and created by practitioners and researchers in the field.

In this paper, we attempted to operationalize the Thurstone scaling through an IRT approach by following a two-step procedure: (1) establish a continuous, or at least an approximate, intensity scale by locating each item on this scale through a sample of experts; and (2) map the individual onto this scale by examining the individual's discrete response (e.g., binary agree/disagree to the statement of the item). The proposed Thurstonian LLTM represents a method for this operationalization. As a method grounded in IRT, the LLTM accordingly inherits many of the advantages of the IRT for scaling multiple dichotomous and polytomous responses.

There are some limitations to the current approach. The Rasch model appears to be too restrictive for capturing the diversity in the data, and the ICCs of the 31 items were not as diverse as we expected. A two-parameter logistic LLTM (e.g., Ip et al. 2009) may be more appropriate. Furthermore, in this paper only item attributes were considered, and person attributes such as experience with the professional school were not taken into account. Currently, further work that expands the Rasch model to its two-parameter logistic counterpart and a regression model incorporating person attributes is in progress.

References

- Allison DB, Basile VC, Yunker HE (1991) The measurement of attitudes toward and beliefs about obese persons. *Int J Eat Disord* 10:599–607
- Babbie ER (2008) *The basics of social research*, 4th edn. Thomson Learning, Inc., Belmont, CA
- Bergner M, Bobbitt RA, Carter WB, Gilson BS (1981) The sickness impact profile: development and final revision of a health status measure. *Med Care* 14:787–805
- Brown A, Maydeu-Olivares A (2012) Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behav Res Methods* 44:1135–1147
- Caffo B, An M, Rohde C (2007) Flexible random intercept model for binary outcomes using mixture of normals. *Comput Stat Data Anal* 51:5220–5235
- Coombs CH (1964) *A theory of data*. Wiley, New York
- Edwards AL, Kenney KC (1946) A comparison of the Thurstone and Likert techniques of attitude scale construction. *J Appl Psychol* 30:72–83
- Fischer GA (1973) The linear logistic test model as an instrument in educational research. *Acta Psychol* 37:359–374
- Ip EH (2010) Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *Br J Math Stat Psychol* 63:395–415
- Ip EH, Smits D, De Boeck P (2009) Locally dependent linear logistic test model with person covariates. *Appl Psychol Meas* 33:555–569
- Ip EH, Marshall S, Crandall SJ, Vitolins M, Davis S, Miller D, Kronner D, Vaden K, Spangler J (2013) Measuring medical student attitudes and beliefs regarding obese patients. *Acad Med* 88:282–289

- Lewis RJ, Cash TF, Jacobi L, Bubb-Lewis C (1997) Prejudice toward fat people: the development and validation of the antifat attitudes test. *Obes Res* 5:297–307
- Likert R (1932) A technique for the measurement of attitudes. *Arch Psychol* 22(140):1–55
- Lord FM (1980) Applications of item response theory to practical testing problems. Erlbaum, Hillsdale, NJ
- Mair P, Hatzinger R (2007) Extended Rasch modeling: the eRm package for application of IRT models in R. *J Stat Softw* 20(9). Last assessed December 11th, 2013. <http://www.jstatsoft.org/v20/i09/paper>
- Massof RW (2002) The measurement of vision disability. *Optom Vis Sci* 79:516–552
- Ostram TM (1989) Interdependence of attitude theory and measurement. In: Pratkanis AR, Breckler SJ, Greenwald AG (eds) *Attitude structure and function*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp 11–36
- Roberts JS, Laughlin JE (1996) A unidimensional item response model for unfolding responses from a graded disagree–agree response scale. *Appl Psychol Meas* 20:231–255
- Thurstone LL (1925) A method of scaling psychological and educational tests. *J Educ Psychol* 16:433–451
- Thurstone LL (1928) Attitudes can be measured. *Am Coll Sociol* 33:529–554
- Thurstone LL, Chave EJ (1929) *The measurement of social attitudes*. University of Chicago Press, Chicago
- Williams VSL, Pommerich M, Thissen D (1998) A comparison of developmental scales based on Thurstone methods and item response theory. *J Educ Meas* 35:93–107
- Zeger SL, Liang KY, Albert P (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44:1049–1060

Quantitative Psychology Research

The 78th Annual Meeting of the Psychometric Society

Millsap, R.E.; Bolt, D.M.; van der Ark, L.A.; Wang, W.C.

(Eds.)

2015, IX, 486 p. 98 illus., 46 illus. in color., Hardcover

ISBN: 978-3-319-07502-0