Investigating Reliability of Fully-labeled and Partially-labeled Rating Scales

An Application of Generalizability Theory

Alfonso J. Martinez[1]

Department of Psychological and Quantitative Sciences
University of Iowa

April 29, 2021

[1]Email: alfonso-martinez@uiowa.edu

Introduction

- Rating scales (e.g., Likert scales) have become an indispensable tool in the social sciences and abroad
- Especially used by researchers/practitioners who are interested in measuring latent (unobserved) constructs
- Rating scales are useful for providing information about these constructs
- Popularity has catalyzed research evaluating scale characteristics and their effect on measurement quality
- Scale design characteristic that has received relatively little attention: differential scale labels (focus of present research)

Motivation

- Scale characteristics → cognitive response processes → measurement/psychometric quality of response data
- Thought experiment:

  I am satisfied with my life.[2]

  ○ 1 = "Agree" ○ 2 = "Neutral" ○ 3 = "Disagree"

  vs.

  I am satisfied with my life.

  ○ 1 = "Agree" ○ ○ 3 = "Disagree"

- Does omission of the intermediate response label have a significant effect on respondent's response patterns and/or the psychometric properties of the scale?

[2] The three-point response format is used for sake of illustration and is not the one utilized in the SWLS.

Review of the Literature

Research in this area has resulted in mixed results

- Huck and Jacko (1974) and Wyatt and Meyers (1987)
  - Partially-labeled scales have lower total score means than fully-labeled scales
- Dixon et al. (1984) and Newstead and Arnold (1989)
  - No differences in total score means as a function of labeling
- Dixon et al. (1984) and Eutsler and Lang (2015)
  - Significantly higher variance in scales labeled only at the endpoints
- Chang (1997)
  - Labeling accounts for $< 1\%$ of variance in observed scores
- Menold et al. (2014)
  - Eye-tracking technology; fixation times were shorter in partially-labeled scales, but each response category received more attention

Present Study

Purpose of the present study

- Explore and compare the reliability and measurement error of rating scales under different scale label configurations
  - To what extent does manipulating the response labels affect the psychometric and measurement properties of rating scales?
- Implement Generalizability theory (G-theory), a flexible statistical framework that integrates analysis of variance techniques with classical test theory (Shavelson et al., 1989)

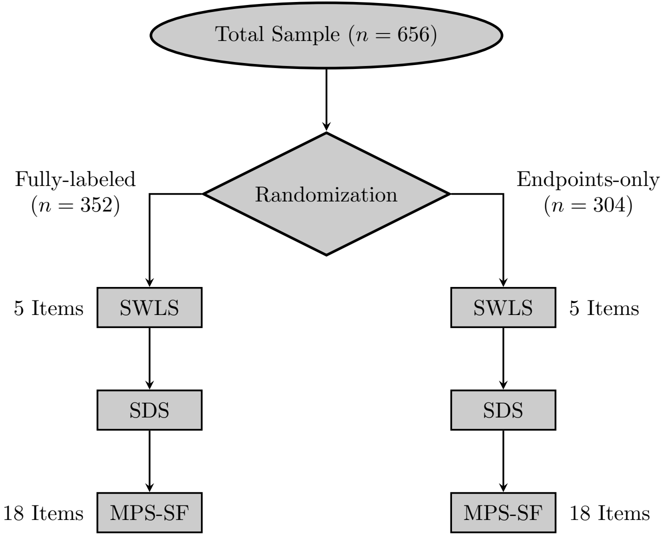
Methods

- $N = 656$ respondents from a university in Central California participated in the study ($M_{\text{age}} = 19.16$, $SD_{\text{age}} = 1.62$)
- Three scales were administered:
  - Satisfaction with Life Scale (SWLS; Diener et al., 1985)
  - Social Desirability Scale (SDS; Crowne & Marlowe, 1960)
  - Multidimensional Perfectionism Scale-Short Form (MPS-SF; Hewitt & Flett, 1990)
- SDS served as distractor task (constant across conditions)
- SWLS and MPS-SF are 7-point rating scales with identical response options
- Two survey sets: Fully-labeled and partially-labeled (endpoints-only)

Scale Labels

SWLS & MPS-SF	Fully-labeled	1 = Strongly Disagree	2 = Disagree	3 = Somewhat Disagree	4 = Neither Agree nor Disagree	5 = Somewhat Agree	6 = Agree	7 = Strongly Agree
	Endpoints-only	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree

Experimental Design



Analysis: Generalizability Theory

G-theory is a statistical framework that "combines" Analysis of Variance and classical test theory (see Brennan, 2001)

- Decomposes an observed score X into the sum of a true score T and k error components
  - $X = T + (E_1 + \cdots + E_k)$ (1)
- Allows user to specify measurement conditions of interest (facets)
- G-theory allows us to estimate amount of variance due to each facet
- Special case of a random effects model
- Rich conceptual framework (G-study and D-study)
  - allows us to how quantify how psychometric properties change as conditions of measurement vary

Our experimental design leads to a $p \times i$ design

$$\text{Model: } X_{pi} = \mu + \mu_i + \mu_p + \mu_{pi}$$

Analysis: Generalizability Theory

The following indices were estimated and compared across conditions

- Absolute error: $\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(pI)$
- Relative error: $\sigma^2(\delta) = \sigma^2(pI)$
- Generalizability coefficient (analogous to coefficient α)
  - $\mathbb{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\delta)}$
- Coefficient of dependability
  - $\Phi = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\Delta)}$
- (Relative) signal-noise ratio:
  - $\mathsf{S/N}(\delta) = \frac{\mathbb{E}\rho^2}{1-\mathbb{E}\rho^2}$
- Models estimated using GENeralized analysis Of VAriance (GENOVA; Brennan, 2001) statistical software program

Results: Measurement Error

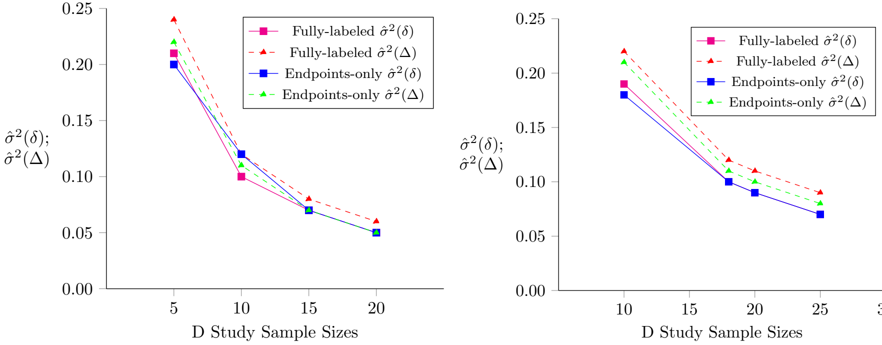


Figure: Absolute and relative error variance estimates (SWLS and MPS-SF)

Results: Reliability - Generalizability Coefficient

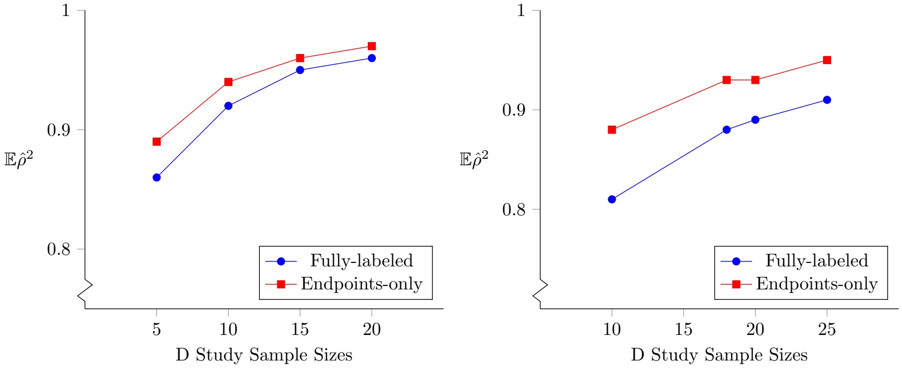


Figure: Generalizability coefficient estimates (SWLS and MPS-SF)

Results: Reliability - Coefficient of Dependability

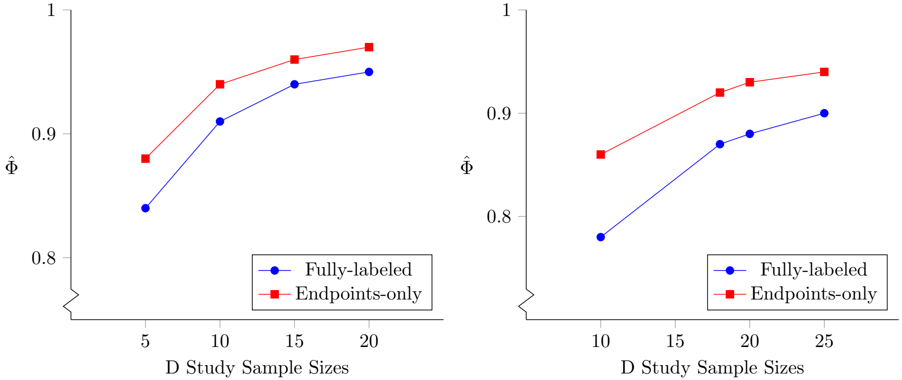


Figure: Coefficient of dependability estimates (SWLS and MPS-SF)

Results: Relative Signal-Noise Ratios

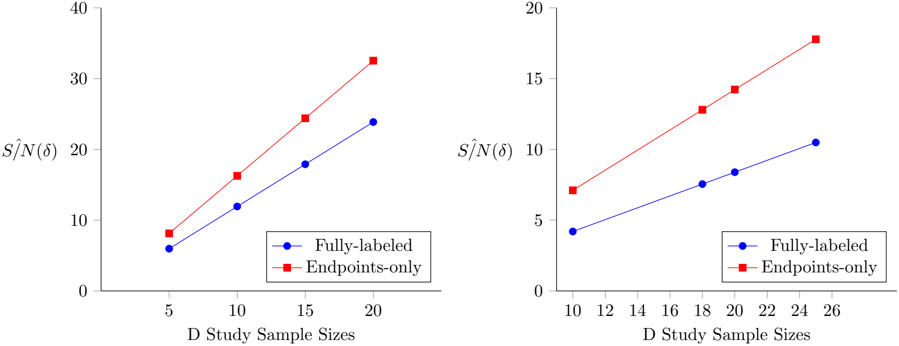


Figure: Relative signal-noise ratio estimates (SWLS and MPS-SF)

Summary of Results

- Endpoint-only scales had higher reliability and signal-noise ratios than their fully-labeled counterparts
- Findings were found on both the SWLS and MPS-SF
- Error variances (absolute and relative) were similar across conditions
- Subsequent item-level analyses (not shown here) indicated that proportion of endpoint endorsements were significantly higher on several items under the endpoints-only labeling scheme
- Taken together, findings suggest non-trivial differences in measurement error and reliability across differentially labeled rating scales

Discussion & Future Directions

Discussion

- Main take-away: Partially-labeled scales had higher reliability and less measurement error than fully-labeled counterparts
- Results went against our initial hypothesis
- Direct relationship to the reliability-validity paradox (Brennan, 2001)

Limitations

- No hypothesis tests in G-theory
- Due to experimental design, our model did not include a label facet
  - Use an alternating treatment design (leads to a $p \times i \times l$ design)

Directions for future research

- Mixed-methods (e.g., interviews + quantitative analyses)
- Directly model response processes with novel measurement models (e.g., IRTrees, multivariate G-theory)
- Examine measurement invariance through MGCFA
- Investigate the reliability-validity paradox in more detail

References I

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag New York.

Chang, L. (1997). Dependability of anchoring labels of likert-type scales. *Educational and Psychological Measurement*, *57*(5), 800–807.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology.. *Journal of Consulting Psychology*, *24*(4), 349–354.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined likert formats. *Educational and Psychological Measurement*, *44*(1), 61–66.

References II

Eutsler, J., & Lang, B. (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting*, *27*(2), 35–51.

Hewitt, P. L., & Flett, G. L. (1990). Perfectionism and depression: A multidimensional analysis. *Journal of Social Behavior and Personality*, *5*(5), 423–438.

Huck, S. W., & Jacko, E. J. (1974). Effect of varying the response format of the alpert-haber achievement anxiety test.. *Journal of Counseling Psychology*, *21*(2), 159.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods*, *26*(1), 21–39.

Newstead, S. E., & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement*, *49*(1), 33–43.

References III

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory.. *American Psychologist*, *44*(6), 922.

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point likert type response scales. *Educational and psychological measurement*, *47*(1), 27–35.

Thank you!