INTERPRETING IRT PARAMETERS: PUTTING PSYCHOLOGICAL MEAT ON THE PSYCHOMETRIC BONE

Anita M. Hubley, Amery D. Wu, & Bruno D. Zumbo

University of British Columbia, Vancouver, BC, Canada

Presented at the Annual Meeting of the American Psychological Association (APA), Toronto, ON, Canada, August 6-9, 2009

Correspondence: Dr. Anita M. Hubley, Dept. of ECPS, The University of British Columbia,
2125 Main Mall, Vancouver, BC, Canada, V6T 1Z4; e-mail: anita.hubley@ubc.ca

INTRODUCTION

Under the item response theory (IRT) class of psychometric models, typically up to three parameters may be estimated to describe people's response patterns: (1) *a*-parameter or item discrimination, which is an item's ability to discriminate among respondents, (2) *b*-parameter or item difficulty*,* which is the threshold value of an item that a respondent's amount of the latent variable must exceed to endorse the item, and (3) *c*-parameter or lower asymptote of the IRT function, which is the probability of a respondent with very little of the latent variable endorsing an item by chance.

Roskam (1985) advocated the importance of understanding the substantive meanings behind IRT parameters for psychological measures. He conjectured that items in personality inventories showed lower discriminating power when formulated in more general and abstract (i.e., less concrete) terms. Zumbo, Pope, Watson, and Hubley (1997) examined this conjecture empirically and showed that it did not hold; they found small to large and significant positive relationships between word/item abstractness and item discrimination with measures of neuroticism and extraversion. In addition, they reported that word/item abstractness showed mixed correlations with item difficulty and low and nonsignificant correlations with the lower asymptote. Other researchers have explored the relationships between IRT parameters and each of item subtlety (vs. obviousness; Zickar & Ury, 2002) and social desirability (Rouse, Finger, & Butcher, 1999; Zickar & Ury, 2002). In the case of social desirability, Zickar and Ury found it to be unrelated to item discrimination and difficulty whereas Rouse et al. argued that it may be related to the lower asymptote.

PURPOSE OF STUDY

The purpose of the present study was to examine the relationship of five different ratings of items (i.e., wording specificity, availability heuristic, emotional comfort, meaning clarity, social desirability) to IRT parameters estimated from responses to the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977). Our goal was to extend previous research that attempts to add psychological meat to the psychometric bone when interpreting IRT parameters.

METHOD

Participants

Two samples were used in this study. IRT parameters were estimated using the responses to the CES-D from a community sample of 600 adults (310 men, 290 women) ages 17 to 87 years (*M* = 44.2, *s* = 12.9). A separate community sample of 31 men and women ages 17 to 83 (*M* = 38.7, *s* = 21.5) rated each item on each of the five variables (e.g., wording specificity) while completing the CES-D.

Measures & Procedure

1) CES-D: The CES-D consists of 20 items. Responses are made using a four-point scale ranging from 0 = 'less than 1 day' to 3 = '5-7 days'. There are four different scoring methods used with the CES-D: ordinal and three binary approaches (presence, persistence, and extreme persistence) (Gelin & Zumbo, 2003). The presence method (i.e., 0 = 'less than 1 day'; 1 = '1-2 days or more') was used here. A principal components analysis (PCA) of the tetrachoric matrix conducted using FACTOR 7.02 (Lorenzo-Seva & Ferrando, 2006) showed the presence of two factors, which was supported by a parallel analysis using marginally bootstrapped samples (PA-MBS; Lattin, Carroll, & Green, 2003). The first factor (eigenvalue = 8.90, variance associated with the 1st factor = 44.5%) consisted of the 16 negatively worded items. The second factor (eigenvalue = 2.96, variance associated with the 2nd factor = 14.8%) consisted of the four positively worded items (#4, 8, 12, 16). In order to have a unidimensional scale for the IRT analyses, the positively worded items were dropped and only the 16 negative items (worded in the depressed direction) were used. A subsequent PCA and PA-MBS showed an essentially unidimensional structure for these 16 items (first eigenvalue = 5.78, variance associated = 36.2%).

2) Item Ratings: A sample of 31 adults independently completed the CES-D items so they could rate each item on the degree to which: (a) the item wording was general vs. specific (wording specificity), (b) their ability to properly respond to the item took a short vs. long time (availability heuristic), (c) they felt uncomfortable vs. comfortable responding to the item (emotional comfort), (d) the meaning of the item was vague vs. clear to them (meaning clarity), and (e) most people would think selecting 'most or all of the time (5-7 days)' as a response to the item would be socially unacceptable vs. acceptable (social desirability). In each case, a 7-point response scale was used. A sample is provided in Figure 1.

3) Rater Characteristics: In addition to their CES-D scores, the 31 raters also provided the following demographic information: age, sex, and educational level.

ANALYSES

Using MULTILOG 7.0, a three-parameter logistic (3PL) model was fit to the data provided by the sample of 600 adults. Given the low average *c*-parameter (*M* = .035, *s* = .061) and the wide range of item discrimination values, we decided to conduct and report the results from a two-parameter logistic (2PL) model. Because one item (#11) had shown a relatively high *c*-parameter of 0.24, we excluded this item from subsequent analyses, which thus used a total of 15 CES-D items.

Next, we correlated the *a*- and *b*-parameters across the 15 CES-D items with each of the five ratings for the items. To be consistent with Zumbo et al. (1997), we used Spearman correlations. Then we obtained the mean and variance of the Spearman correlations across the 31 raters for each item. One-sample t-tests were then computed to determine if the mean correlations were statistically significantly different from zero.

Finally, we conducted a regression analysis to determine to what degree the variance in the Spearman correlations (between the *a*- and *b*-parameters and each of the five item ratings) could be explained by the raters' personal characteristics of age, sex, educational level, or their average CES-D item score.

RESULTS

First, the results of the Spearman correlations of the *a*- and *b*-parameters across the 15 CES-D items with each of the five ratings for the items are presented in Table 1.

- There was a small tendency for items that were rated as (a) being more specific (than general) in their wording and (b) having meanings that were more clear (than vague) to be more difficult (i.e., have higher *b*-parameter values and require more of the latent variable 'depressive symptomatology' for them to be endorsed).

- There was a small tendency for items that had higher social desirability ratings to be less difficult (i.e., have lower *b*-parameter values and require less of the latent variable 'depressive symptomatology' for them to be endorsed) and less able to discriminate among levels of depressive symptomatology (i.e., have lower *a*-parameter values).

- Neither the availability of response nor the emotional comfort level ratings were related to the discrimination or difficulty parameters.

Second, the results of the regression analysis to determine if the Spearman correlations between the *a*- and *b*-parameters and each of the five item ratings could be explained by the raters' personal characteristics showed that none of the personal variables served an explanatory role, with only one exception. The exception was that the correlation between the *a*-parameter and wording specificity was explained by educational level, $\beta = 0.44$, $p < .05$, such that the greater the educational level, the higher the correlation between the *a*-parameter and wording specificity.

DISCUSSION

The present findings contribute to the rather small literature that attempts to provide further psychological meaning to the interpretation of IRT parameters, particularly in the case of non-achievement measures. Most of the item ratings included here (e.g., availability heuristic, emotional comfort, meaning clarity) are new to the literature but only ratings of wording specificity, meaning clarity, and social desirability showed any significant relationship to the *a* and *b* parameters. Notably, the results for social desirability differ from previous research. Zickar and Ury's (2002) reported near-zero correlations between social desirability and both the *a*- and *b*-parameters obtained with a measure of personality whereas, in the present study, small but significant negative correlations were found between social desirability and both parameters obtained with a depression measure. Overall, the present findings make sense within the context of a construct like depressive symptomatology. The differences in findings suggest that the relationship between item ratings and IRT parameters may be more specific to the construct of interest than previously considered.

REFERENCES

Gelin, & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement, 63*, 65-74.

Lattin, J., Carroll, D.J., & Green, P.E. (2003). *Analyzing multivariate data* (pp. 114-116). Belmont, CA: Duxbury Press.

Lorenzo-Seva, U., & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*, 88-91.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *3*, 385-401.

Roskam, E. E. (1985). Current issues in item response theory: Beyond psychometrics. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 3-19). Amsterdam: Elsevier Science.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*, 282-307.

Zickar, M. J. & Ury, K. L. (2002). Developing an interpretation of item parameters for personality items: Content correlates of parameter estimates. *Educational and Psychological Measurement, 62*, 19-31.

Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement*, *57*, 963-969.

Figure 1
Instructions for Rating the CES-D Items

Rating Instruction

You have two tasks. In the first task, please respond to the statement by circling a number on the right that best describes how often you felt or behaved this way during the past week.

1 = Rarely or none of the time (less than 1 day)
2 = Some or a little of the time (1-2 days)
3 = Occasionally or a moderate amount of time (3-4 days)
4 = Most or all of the time (5-7 days)

It is important to respond to each statement so that you may better complete the second task.

In the second task, please rate the statement by circling a number for each of the five rating descriptions below it. You will be asked to rate 20 statements in total. For each statement, you will complete the same set of 5 rating descriptions.

Meaning of numbers in rating descriptions:

3 = very
2 = somewhat
1 = slightly / a little
0 = neutral

Example

1. My mood often goes up and down. 1 2 3 4

- The wording of this statement was **general** 3 2 1 0 1 2 3 **specific**.
- To answer properly, I had to think a **short** 3 2 1 0 1 2 3 **long** time.
- I felt **uncomfortable** 3 2 1 0 1 2 3 **comfortable** responding to this statement.
- This statement was **vague** 3 2 1 0 1 2 3 **clear** to me.
- Most people would think responding 4 to this statement is *socially* **unacceptable** 3 2 1 0 1 2 3 **acceptable**.

Note: please respond to the last rating task, even if you did not circle "4".

Table 1
Relationships between 2PL Model IRT Parameters and CES-D Item Ratings

	Wording Specificity	Availability Heuristic	Emotional Comfort	Meaning Clarity	Social Desirability
a-parameter (discrimination)	.05	-.05	-.06	.03	-.15
	n.s.	*n.s.*	*n.s.*	*n.s.*	*p* < .01
b-parameter (difficulty)	.11	-.05	.003	.12	-.11
	p < .05	*n.s.*	*n.s.*	*p* < .05	*p* < .05