Psychometric Study of the CES-D: Factor Analysis and DIF

Bruno D. Zumbo, Michaela N. Gelin, & Anita M. Hubley
University of British Columbia

Presented at: ***International Neuropsychological Society's 29th Annual Meeting, Chicago IL February 16, 2001.***

Abstract: We report on a psychometric study of the Center for Epidemiologic Studies Depression (CES-D) scale with 600 community-dwelling adults between the ages of 17 and 87 years. The mean age for males is 46 years (N=310) and 42 years for females (N=290). We propose and test a one-factor measurement model with confirmatory factor analysis, which takes into account method effects. The method effects represent the distinction between positively and negatively worded items. Also, we studied gender based differential item functioning (DIF) using a method proposed by Zumbo (1999). These DIF analyses were followed-up by nonparametric item response (IRT) DIF and differential test functioning. Our results indicate that the proposed measurement model fits and hence helps one understand the disparate literature on the factorial structure of the CES-D. This one factor model was also completely invariant (including method effects) across genders. With regard to the item level analyses investigating the DIF, "crying" and "eating" displayed gender DIF. This item-level DIF translates to substantial effects in scale score interpretation. This is the first study on the CES-D to have modeled the method effects in a one-factor measurement model, tested these method effects across genders, and to have examined gender DIF using Zumbo's method.

Send correspondence to: Dr. Bruno D. Zumbo
University of British Columbia
Scarfe Building, 2125 Main Mall
Department of ECPS
(Program Area: Measurement, Evaluation, & Research Methodology)
Vancouver, B.C.
CANADA V6T 1Z4

e-mail: bruno.zumbo@ubc.ca

web page: http://www.educ.ubc.ca/faculty/zumbo/

- The CES-D (Radloff, 1977) is a widely used self-report measure developed for use in studies exploring the epidemiology of depressive symptomatology in the general population.
- Few studies have examined both the item and scale level psychometric properties of the CES-D with a large community-dwelling sample.
- The scale has been used in numerous studies to: (a) compare the prevalence of depressive symptomatology between men and women, (b) select a non-depressive sample for a research study, or (c) split a sample into depressed and non-depressed groups for comparison on some other variable of interest.
- In all of these cases it is important that (a) the scale performs in the same manner it is scored – it is summed to one score hence a factor analysis should result in one-factor, and (b) that this one-factor solution is invariant across genders. If the measure is not invariant across genders then one gender will artificially be portrayed as having a higher prevalence of depression, and group differences based on the depression score may be confounding gender effects.
- Community-dwelling sample was obtained in a survey with the Institute for Social Research and Evaluation at the University of Northern British Columbia :

Report

Your present age?

Sex	Mean	N	Std. Deviation	Minimum	Maximum
female	42.19	290	13.44	18	87
male	46.05	310	12.07	17	82
Total	44.19	600	12.88	17	87

- Gender differences in Depressive Symptomotology:

CESD

Sex	Mean	N	Std. Deviation	Minimum	Maximum
female	10.9	290	9.4	.0	46.0
male	9.4	310	9.4	.0	50.0
Total	10.1	600	9.5	.0	50.0

Conducting a t-test, t(598)=1.89, p=0.059 (n.s.)

- Applying cut-score of 15/16 for depression we can explore gender differences.

Crosstabulation of Sex by Depression

			DEPRESSED		
			NO	YES	Total
Sex	female	Count	211	79	290
		% within Sex	72.8%	27.2%	100.0%
	male	Count	261	49	310
		% within Sex	84.2%	15.8%	100.0%
Total		Count	472	128	600
		% within Sex	78.7%	21.3%	100.0%

- A Chi-squared test indicates a statistically significantly higher proportion of depressed females than males. Chi-squared with 1 df equal 11.6, p=0.001
- The question, however, is whether this difference (which has been found in several other studies) is a measurement artifact.
    - Two methods used to investigate the possible measurement artifact (a) scale level analysis – confirmatory factor analysis, (b) item level analysis – differential item functioning
    - First, however, a confirmatory factor analysis is needed for the whole sample, irrespective of sex.
- Two models (see Figures 1 and 2) were tested with LISREL 8.30 using: Given the item rating format a polychoric correlation, generally weighted least-squares estimation with the asymptotic covariance matrix was used so that the correct standard errors are produced.
- Given that 4 of the CES-D items are worded in a positive manner (16 others are negative wording) we postulated a method effect due to item wording. We were able to incorporate this method effect by modeling certain correlated uniquenesses among the 4 items, over-and-above their loadings on the common factor.
- Furthermore, we computed a reliability-like estimate with and without the method effects to see how the reliability is affected.

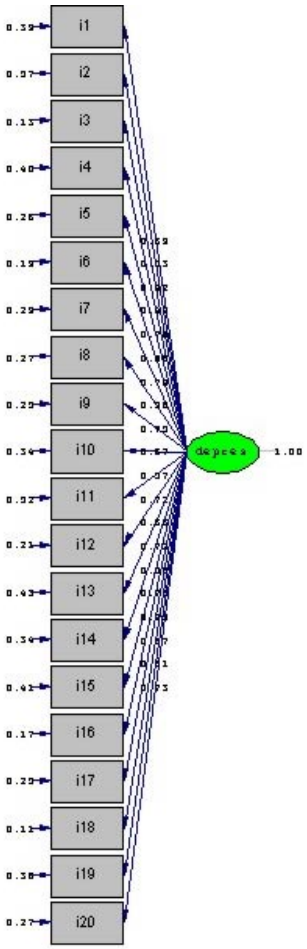
$$\text{Scale reliability} = \frac{\sum(\text{standardized loadings})^2}{\sum(\text{standardized loadings})^2 + \sum(\text{squared standardized errors})}$$

- Model 2 (allowing for method effect due to item wording) fits well and is statistically better fitting than Model 1, (i.e., testing the nested models: Chi-squared df=6 of 205.2, p=0.000).
- As predicted by theory and simulation (Zimmerman, Zumbo, & Lalonde, 1993) the reliability for Model 1 (i.e., when ignoring the correlated errors) will be inflated by the number of correlated errors and their magnitude. The six error correlations among the four positively worded items ranged from 0.21 to 0.37. Coefficient alpha for the scale is 0.91.
- Model 2 was fit for males and females using a simultaneous multi-group confirmatory factor analysis (LISREL 8.30) using the same methods as described above. Note that this across group fit was for complete (i.e., strict) invariance: loadings, error variance, and method effect correlations between the same across the two genders.

- The resulting RMSEA=0.044 with a test of close fit indicates of a fully invariant model. In short, the scale level results indicate that we are measuring the same thing in both males and females, to the same level of precision, and with the same method effects.
- The positively worded items are introducing extra covariation in people's responses; this extra covariation needs to be taken into account in scale level analyses such as factor analysis.
- Next, an item analysis was computed to investigate whether item level differential item functioning is present for males and females. As recently shown by Zumbo (in press) item-level DIF will not necessarily manifest itself in scale level analyses such as factor analysis.
- On an item-by-item basis, differential item functioning is present if, after conditioning/matching on the variable of interest (in our case depressive symptomotology), the groups of interest (in our case males and females) differ statistically on the item score (Swaminathan, 1994; Swaminathan & Rogers, 1990; Clauser & Mazor, 1998).
- Because the items of the CES-D are Likert-type, we used Zumbo's ordinal logistic regression method to detect differential item functioning. This allows for both uniform (main effect of group) and non-uniform DIF (interaction of group differences by conditioning variable).
- No items showed non-uniform DIF. Only the "eating" and "crying" items (#2 and #17) showed statistically significant uniform gender DIF. Females were 2.24 times more likely to respond with a higher item score on the "eating" item than males. Females were also 9.3 times more likely to respond with a higher item score on the "crying" item than males – scoring in a more depressive fashion.
- This indicates that even if one matches males and females on their overall level of depressive symptomotology, females are still more likely to respond in what is considered a depressive manner for these two items. Given that one has matched on depressive symptomotology, this is suggestive of a measurement artifact.
- We conducted a non-parametric item response theory analysis (TestGraf: Ramsey, 2000, 1991) to understand the DIF. (Group 1= males and Group 2 = females)
- The reader should cautiously interpret the item curves at the upper end of the depressive symptomotology scale (particularly scores above 30) because there are few respondents with that level of depression so the curves have a relatively large amount of sampling variability (i.e., there is not much information or data from which to plot the curves that high on the continuum).
- Therefore, focusing on the 0 to 30 score range, we can see from Figure 3 that for both items #2 and #17 females (Group 2) tend to score higher even though they are matched on overall depressive symptomotology. They are matched by having their item response functions plotted on the same scale. This supports the results from Zumbo's method above.

- Figure 4 provides for us the differential test functioning (i.e., how does the item bias effect the scale level score). Note that the dashed lines are the various percentile scores. Therefore, the second horizontal dashed line from the top indicates the 75th percentile for female respondents (in this case a score of 16). If one follows this horizontal dashed line across to the plotted curve and then follows the vertical dashed line at that point to the scores for a match male respondent, one sees that this corresponds to a score for males of approximately 12. Hence there appears to be nearly a 4-point difference between matched males and females near the cut-score of the CES-D.
- The differential item and test functioning (even though the CFA showed strict measurement invariance) is a concern that needs further study by depression researchers to investigate whether one would want to remove these items from the scale. Also, the differential item and test functioning may explain why females are consistently being seen as more depressed than males on the CES-D.

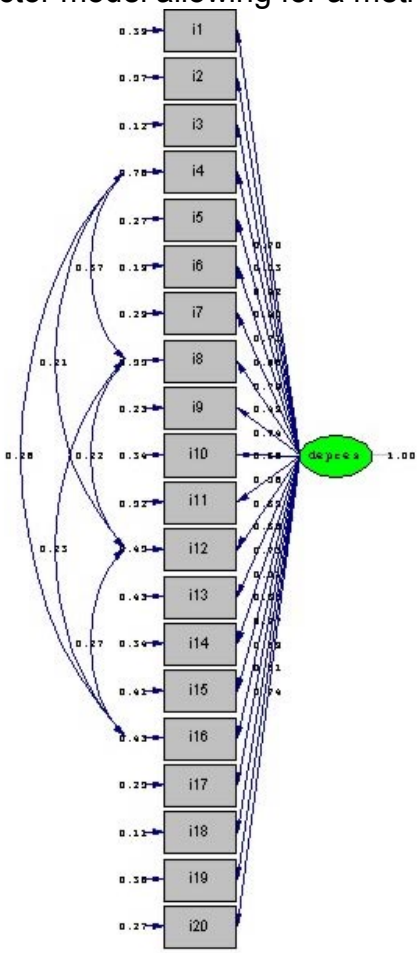
Figure 1. Model 1: Single-factor model



Root mean square error of approximation (RMSEA)=0.086
P-value for the test of close fit (RMSEA <0.05)= 0.00 ∴ Not an adequate fit of the model.

Reliability of 0.968

Figure 2. Model 2: Single-factor model allowing for a method effect



Root mean square error of approximation (RMSEA)=0.051
P-value for the test of close fit (RMSEA <0.05)= 0.33 ∴ Adequate fit of the model.

Reliability of 0.962

Figure 3. Nonparametric DIF Plots for the Two DIF Items

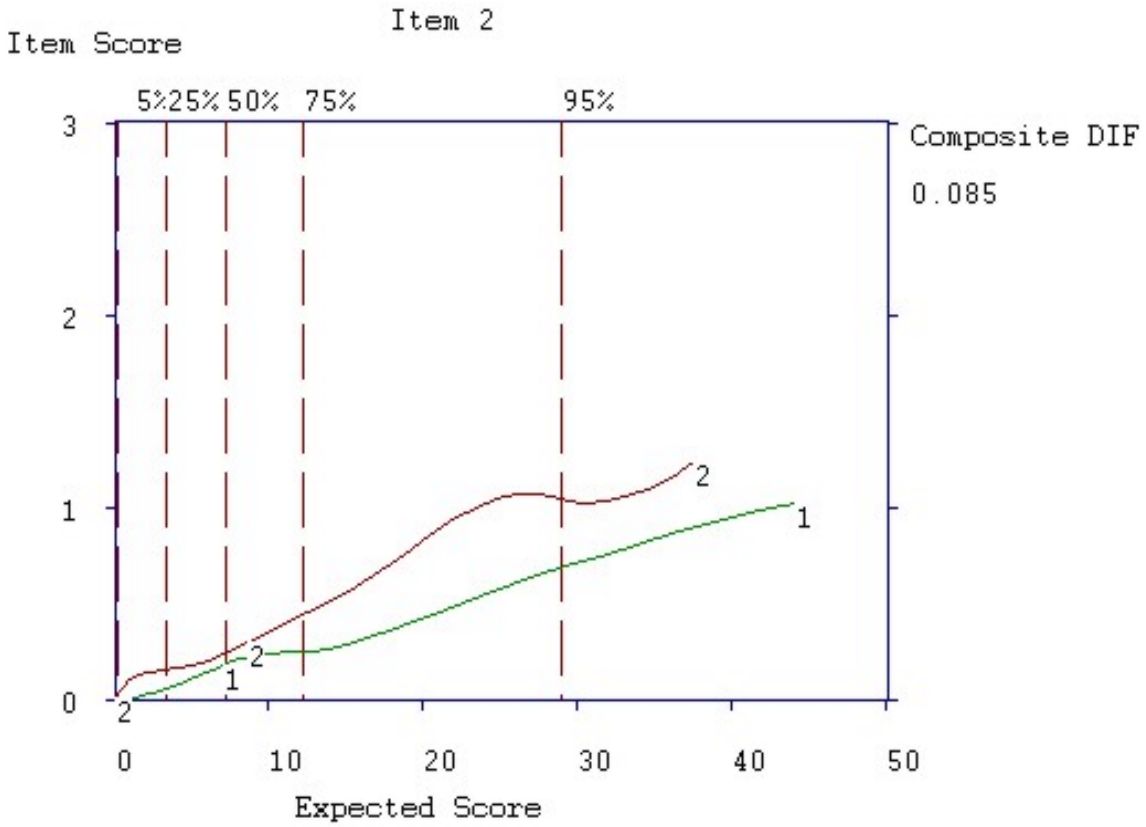
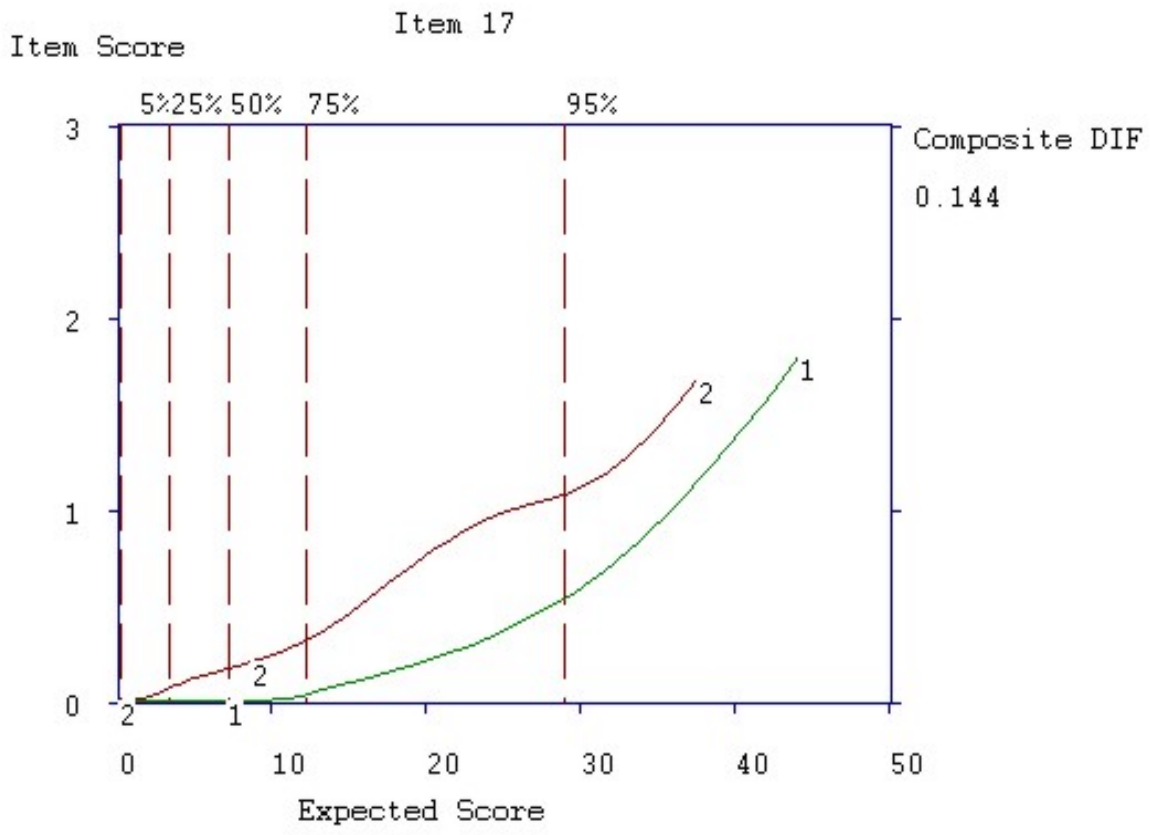
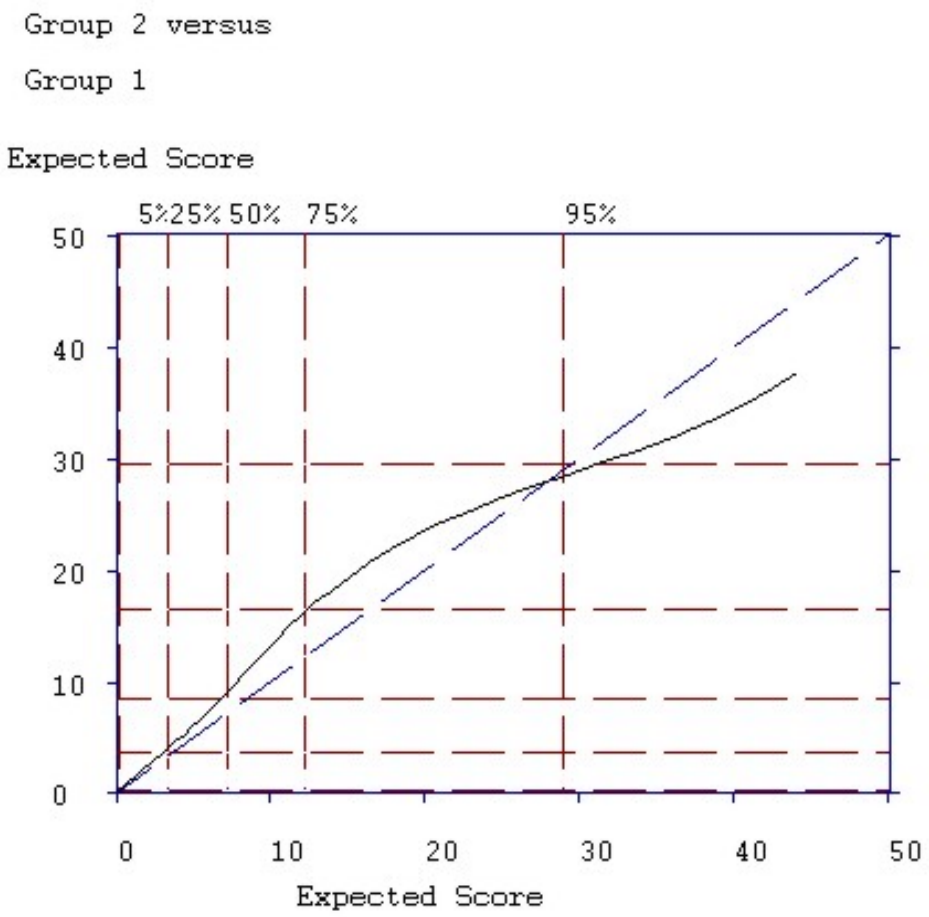


Figure 3 (cont'd.).



Note: Because the scale scores are clustered toward the lower end of the scale, the curves above the 95th percentile should not be considered in the conclusions (too little statistical information at that level of the continuum).

Figure 4. Nonparametric Differential Test Functioning Plot.



Note: The vertical axis = scores by females; horizontal axis = scores by males. Also, because the scale scores are clustered toward the lower end of the scale, the curve above the 95th percentile should not be considered in the conclusions (too little statistical information at that level of the continuum).

The CES-D Items.

For each statement, circle the number (see the guide below) to indicate how often you felt or behaved this way **during the past week**.

0 = rarely or none of the time (less than 1 day)
1 = some or a little of the time (1-2 days)
2 = occasionally or a moderate amount of time (3-4 days)
3 = most or all of the time (5-7 days)

	not even 1 day	1-2 days	3-4 days	5-7 days
1. I was bothered by things that usually don't bother me.	0	1	2	3
2. I did not feel like eating; my appetite was poor.	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends.	0	1	2	3
4. I felt that I was just as good as other people.	0	1	2	3
5. I had trouble keeping my mind on what I was doing.	0	1	2	3
6. I felt depressed.	0	1	2	3
7. I felt that everything I did was an effort.	0	1	2	3
8. I felt hopeful about the future.	0	1	2	3
9. I thought my life had been a failure.	0	1	2	3
10. I felt fearful.	0	1	2	3
11. My sleep was restless.	0	1	2	3
12. I was happy.	0	1	2	3
13. I talked less than usual.	0	1	2	3
14. I felt lonely.	0	1	2	3
15. People were unfriendly.	0	1	2	3
16. I enjoyed life.	0	1	2	3
17. I had crying spells.	0	1	2	3
18. I felt sad.	0	1	2	3
19. I felt that people dislike me.	0	1	2	3
20. I could not get "going".	0	1	2	3

Note: Items 4, 8, 12, and 16 were reverse coded.

References:

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice,17*, 31-44.

Radloff, L. S. (1977). The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Ramsay, J. O. (1991) Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611-630.

Ramsey, J. O. (2000). *TestGraf* [Computer Software], Author, McGill University, Montreal.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 2*7, 361-370.

Swaminathan, H. (1994). Differential item functioning: A discussion. In Dany Laveault, Bruno D. Zumbo, Marc E. Gessaroli, and Marvin W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issue*s. Ottawa, Canada: University of Ottawa.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational & Psychological Measurement, 53*, 33-49.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (in press). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing*.