Construct Validity: An Illustration of Examining Validity Evidence Based on Relationships to Other Variables Using Correlation, Multiple Regression and Discriminant Function Analysis[1]

Robert K. Gable
Director, Center for Research and Evaluation
Alan Shawn Feinstein Graduate School
Johnson & Wales University

[1]Paper presented at the 45th annual meeting of the New England Educational Research Organization, April 18, 2013, Portsmouth, NH

Permissions

Adapted versions of material in this paper will appear in the 3rd ed. of the text entitled *Instrument Development in the Affective Domain* (McCoach, Gable, & Madura, in press) with permission of the publisher, WPS, as follows:

Abstract

This paper discusses an aspect of construct validity referenced in the *Standards* (1999) as "evidence based on relations to other variables". Data from 150 mothers between 2 and 12 weeks postpartum were gathered using the *Postpartum Depression Screening Scale* (*PDSS)*, the *Edinburgh Postnatal Depression Scale*, and the *Beck Depression Inventory-II.* Data from correlations, multiple regression, and discriminant function analysis are illustrated to examine the construct validity of the *PDSS* score interpretations. The procedures employed should be useful to researchers developing affective instruments.

As stated in the *Standards for Educational and Psychological Testing* (1999), there are several sources of validity evidence needed to support the proposed interpretations of scores (p. 11). Noting that “validity” is not a unitary concept, the *Standards* identify sources of evidence that the developer must consider (p. 11). Of particular importance for affective instrument developers are types of evidence based on:

1. test content ,
2. internal structure, and
3. relationships to other variables (p. 13).

This paper will assume that appropriate judgmental and empirical evidence have been developed for points 1 and 2 above, and will concentrate on illustrating “evidence based on relations to other variables”. This additional evidence is needed to support the developer’s argument that the derived scores

actually reflect the targeted constructs.

Example 1: Correlations

Simple correlations are often used to examine the relationship of the obtained scores with scores from a known instrument. This is referred to as developing *convergent evidence* that the scores on the new instrument relate in theoretically correct magnitude and direction with scores from known instruments.

Instrumentation. The *Postpartum Depression Screening Scale* - *PDSS* (Beck & Gable, 2000; Beck & Gable, 2002) provides an example of this correlational technique. The *PDSS* contains 35 items; all are negative feelings to describe how a mother may feel after the birth of her baby (e.g., I felt really overwhelmed; I felt like I was losing my mind). Mothers describe their degree of disagreement or agreement with each statement on a 5-point Likert response scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Thus, higher scores indicate higher levels of postpartum depression. Each of the seven dimensions (i.e., Sleeping/Eating Disturbances, Anxiety/Insecurity, Emotional Lability, Cognitive Impairment, Loss of Self, and Suicidal Thoughts) contains five items. The *PDSS* is designed to measure a mood disorder, postpartum depression, which mothers may experience any time during the first year after delivery of the child. All of the PDSS stems originated from actual quotes from women who had participated in Beck's (1992, 1993, 1996) qualitative research studies of postpartum depression.

Sample. In developing the *PDSS,* the authors identified a group of 150 mothers who were between 2 and 12 weeks postpartum and had no history of diagnosable depression during the pregnancy. This diagnostic sample completed three self-report depression inventories in random order: the *PDSS*, the *Edinburgh Postnatal Depression Scale – EPDS* (Cox et al., 1987), and the *Beck Depression Inventory-II - BDI-II* (Beck, Steer, & Brown, 1996). Each mother then participated in a structured interview using the *Structured Clinical Interview for DSM-IV Axis I Disorders* - SCID (Spitzer, Endicott, & Robins, 1978). Using the *SCID* interview, the researchers classified the mothers into one of three groups (i.e., **No Depression**, *N* = 104; **Depressive Disorder**, *N* = 28, and **Major Depression**, *N* = 18). Beck and Gable (2000, 2002) examined "test-criterion relationships" (i.e., the criterion was postpartum depression group membership) by computing correlations among the *PDSS* Total score, total scores from two other well-known self-report depression inventories (the *EPDS* and the *BDI-II*), and depression diagnostic status (no depression: *N* = 104 versus a combined depressive disorder and major depression group: *N* = 46) as derived from the SCID interview.

Findings - Correlations. The *PDSS* Total score was strongly correlated ($p < .001$) with *BDI-II* score ($r = .81$, $r^2 = .66$), *EPDS* score ($r = .79$, $r^2 = .44$), and *SCID* diagnostic status (i.e., a combined depression disorder/major depression group vs. no depression; r = .70, r^2 = .49). Thus the *PDSS* was highly associated with both other established self-report depression inventories, as well as depression status as ascertained by a clinical interview. Collecting evidence of convergent validity using multiple variables provided validity evidence for the *PDSS* score interpretations (Beck & Gable, 2002).

Example 2: ***Multiple Regression***

Beck and Gable (2001) also demonstrated how a correlational approach using multiple regression can be employed to examine "test-criterion relationships" to provide evidence to support the validity of construct interpretations. In this example Beck and Gable employed the concept of ***incremental validity*** to determine if the *PDSS* would predict the criterion (group membership: depressed vs. non-depressed) over and above traditionally used instruments. Since the *PDSS* total score was shown to be highly correlated to the SCID depression diagnostic status (r = .81), the authors examined the extent that the *PDSS* total score contributed incrementally to prediction of variance in the criterion (i.e., diagnostic status), above and beyond the variance explained by the best instruments currently available (Cronbach, 1971; Cronbach & Gleser, 1957; Cronbach & Meehl, 1957. Hierarchical regression analyses examined the extent that the *PDSS* could increment the explanation of variance in the *SCID* diagnostic status controlling for the *BDI-II* and *EPDS*. Beck and Gable (2002) reported the following:

Findings - Regression. The results of the regression analysis are presented in Table 1 (attached). The *BDI-II, EPDS, and PDSS* were entered sequentially into the equation. The criterion variable was SCID diagnostic status (classification of the women into depressed or non-depressed groups). The amount of variance in the criterion explained by each predictor is listed in the column labeled "Increase R^2." The data demonstrates that all three depression questionnaires account for a significant proportion of the variance in SCID diagnostic status, as would be expected given the strong correlations among these variables. Entered first, the *BDI-II* accounted for 38% of the variance (p < .001) in group classification. The *EPDS* accounted for an additional 3% *(p* < .05) of the variance. Entered last, the *PDSS* explained an additional 9% (p < .001) of the variance in depression diagnosis. These results show that the *PDSS* offers additional power to predict the assignment of women to depressed or non-depressed groups, even after the predictive abilities of the *BDI-II* and *EPDS* have been statistically removed. This increase in prediction of group classification (i.e., incremental validity evidence) provides further support for the postpartum depression construct assessed by the *PDSS* (Beck & Gable, 2002, p. 42). The logical reason for the successful finding for the incremental validity of the *PDSS* scores was that the other two well-known depression measures assessed "general depression" attributes and the *PDSS* assesses attributes specific to postpartum depression.

Example 3: *Discriminant Function Analysis*

Continuing with the theme of **validity evidence based on relations to other variables** (see earlier p. 1 of this paper), known group differences can be examined to provide concurrent validity evidence of "instrument - criterion" relationships. To illustrate the use of discriminant function analysis (DFA) for developing this validity evidence, Beck and Gable (2001, 2002) also examined the accuracy of using the *PDSS* total score to assign mothers to the three externally determined SCID (i.e., Structured Clinical Interview for DSM-IV Axis I Disorders) diagnosed groups identified previously: **no depression, depressive disorder and major depression.** This type of known groups analysis investigates whether the new instrument could successfully distinguish among groups that had been professionally diagnosed. Successful classification provides evidence for the construct validity of the *PDSS* score interpretations. As previously noted, the first group consisted of **104** women who did not receive a SCID depression diagnosis, the second group comprised **28** women who were diagnosed with Depressive Disorder Not Otherwise Specified (NOS; i.e., *DSM-IV* terminology for significant symptoms of depression that are not severe enough to meet criteria for Major Depressive Disorder), and the third group consisted of **18** women who were diagnosed with Major Depressive Disorder with Postpartum Onset.

Findings - DFA. Table 2 presents the results of the discriminant function analysis; the two canonical discriminant functions were significant predictors of diagnostic group membership. Overall, the procedure correctly classified 115 women (76.7% of the diagnostic sample). The accuracy rates varied across groups: The DFA correctly classified 88 of 104 women (84.6%) with no depressive diagnosis, 15 of 28 women (53.6%) with Depressive Disorder NOS, and 12 of 18 (66.7%) women with major postpartum depression. All 18 women diagnosed with major depression were classified in one of the two depression groups. This means that the *PDSS* discriminant function classification procedure yielded no false negatives for women with the most severe depression diagnosis, which is a highly desirable characteristic for a screening instrument (Beck & Gable, 2002).

Table 3 presents the correlations between scores on the seven *PDSS* content scales and the first canonical discriminant function (which accounted for 92% of the variance explained by the two functions). These correlations assess the relative contributions of the content scales to the classification results. As Table 3 shows, all seven content scales were substantial predictors of classification. The Anxiety/Insecurity scale explained the most variance in classification and the Suicidal Thoughts scale explained the least variance in classification. The diagnostic sample was fairly homogenous on the Suicidal Thoughts variable, which may help to explain why the coefficient for that variable was relatively low (Beck & Gable, 2002).

Summary

This paper has discussed how “evidence based on relations to other variables” provides evidence needed to support an instrument developer’s argument that the derived score actually reflects the targeted constructs. The uses of correlations, multiple regression and discriminant function analysis to provide the validity evidence were illustrated.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing.* Washington: AERA.

Beck, C. T., (1992). The lived experience of postpartum depression: A phenomenological study. *Nursing Research, 41,*166-170.

Beck, C. T., (1993). Teetering on the edge: A substantive theory of postpartum depression. *Nursing Research, 42,* 42-48.

Beck, C. T. (1996). Postpartum depressed mothers' experiences interacting with their children. *Nursing Research, 45,* 98-104.

Beck, C. T., & Gable, R. K. (2000). *Postpartum Depression Screening Scale*: Development and psychometric testing. *Nursing Research, 49*, 272-282.

Beck, C. T., & Gable, R. K. (2001). Further validation of the *Postpartum Depression Screening Scale*. *Nursing Research, 50,* 155-164.

Beck , C. T., & Gable, R. K. (2002). *Technical Manual: Postpartum Depression Screening Scale*. Los Angeles: Western Psychological Services.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II manual*. San Antonio: The Psychological Corporation.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd Ed.), Washington, DC: American Council on Education.

Cronbach, L. J., & Glessor, G. C. (1957). *Psychological test and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Cox, J. L., Holden, J. M., & Sagovsky, R. (1987). Detection of postnatal depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *British Journal of Psychiatry,150,* 782-786.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry*, *35,* 773-782.

Table 1

Incremental Prediction of SCID Depression Diagnosis
by Self-Report Inventories

Predictors	R^2	Increase R^2	Beta
BDI-II[a]	.38	.38**	.15
EPDS[b]	.41	.03*	.01
PDSS[c]	.50	.09**	.56

Note. Hierarchical regression analysis with SCID depression diagnostic status as dependent variable. Diagnostic sample (N = 150).

[a] Beck Depression Inventory, 2nd ed., total score.

[b] Edinburgh Postnatal Depression Scale, total score.

[c] Total score.
*p<.05
**p .001

Table 2

Discriminant Function Results for the Postpartum Depression Screening Scale

Actual Group	Predicted Group Membership		
	No (1)	DD(2)	Major (3)
Group 1	**88**	16	0
No Depression (N = 104)	(84.6%)	(15.4%)	(0%)
Group 2	8	**15**	5
Depressive Disorder NOS (N = 28)	(28.6%)	(53.6%)	(17.9%)
Group 3	0	6	**12**
Major Depression (N = 18)	(0%)	(33.3%)	(66.7%)

Note. Diagnostic sample (*N* = 150). Procedure correctly classified 76.7% of original cases, a rate which is 24% above that from chance alone. Predicted group membership variable is based on the SCID depression diagnosis.

Table 3

Correlations between PDSS Symptom Content Scales and Canonical Discriminant Function

PDSS Content Scale	Correlation with First Discriminant Function
Sleeping/Eating Disturbances	.63
Anxiety/Insecurity	.82
Emotional Lability	.70
Mental Confusion	.64
Loss of Self	.66
Guilt/Shame	.64
Suicidal Thoughts	.48