DOCUMENT RESUME

ED 359 198 TM 019 819

AUTHOR Crowley, Susan L.; And Others
TITLE Depression in Children: The Children's Depression Inventory.
PUB DATE 25 Mar 93
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 11-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Children; *Clinical Diagnosis; *Depression (Psychoicgy); Diagnostic Tests; Emotional Disturbances; Estimation (Mathematics); *Generalizability Theory; Measurement Techniques; *Measures (Individuals); Personality Problems; Scores; Testing Problems; *Test Reliability; Test Theory
IDENTIFIERS *Childrens Depression Inventory; Texas

ABSTRACT

Issues surrounding accurate assessment of depression in children have received much attention. However, the stability of scores from depression measures has generally been estimated using only classical test score theory, rather than the more powerful generalizability theory. The dependability of scores from the Children's Depression Inventory (CDI) was investigated using both generalizability and classical test score analyses. Data from 164 children aged 11 to 16 years from small Texas communities were the basis for the analysis. Results suggest that the sources of error variance interact to decrease the dependability of CDI scores. Several sample measurement protocols were also investigated. The results indicate that depression in children might be better assessed using planned multiple testing sessions to reflect depression more accurately. Two tables present study findings, and there is an appendix of selected study results. (Author/SLD)

aeracdi.wp1 3/25/93

Depression in Children: The Children's Depression Inventory

Susan L. Crowley

Utah State University 84322-2810

Bruce Thompson

Texas A&M University
and
Baylor College of Medicine

Frances Worchel

Texas A&M University

Paper presented at the annual meeting (session #56.31) of the American Educational Research Association, Atlanta, GA, April 16, 1993.

ABSTRACT

Issues surrounding accurate assessment of depression in children have received much attention. However, the stability of scores from depression measures has generally been estimated using only classical test score theory, rather than the more powerful generalizability theory. The present study investigated the dependability of scores from the Children's Depression Inventory using both generalizability and classical test score analyses. Results suggest that the sources of error variance interact to decrease the dependability of CDI scores. Several sample measurement protocols were also investigated. Results indicate that depression in children might be better assessed using planned, multiple testing sessions.

Depression in Children:

The Children's Depression Inventory

Depression in children can have a profound negative impact on self-esteem, peer-relationships, and educational achievement. Depressive symptoms have been related to poor academic achievement, peer relationship problems, behavior problems, poor self-esteem, and in severe cases, suicide (Carlson & Cantwell, 1980; Kazdin, 1990; Worchel, Nolan, & Willson, 1987). Consequently, practitioners and researchers have sought to assess depressive symptomatology to effectively identify children seriously in need of treatment and those evidencing milder depressive symptoms who may nevertheless benefit from early intervention.

Considerable effort has gone toward the development of instruments that yield valid data regarding depressive symptoms. Instruments have been developed for administration to teachers, parents, peers, and the children themselves (self-report). Evidence regarding the reliability of the scores from these instruments, derived using classical test theory, has been widely reported in the literature.

However, more recently, generalizability theory has been proposed as a broader and more powerful model for estimating the dependability of scores from behavioral measurements (Cronbach, Gleser, Nanda, & Rajaratnum, 1972). As Jaeger (1991) notes, given the availability of this newer measurement theory,

Thousands of social science researchers will no

> longer be forced to rely on *outmoded* [classical theory] reliability estimation procedures when investigating the consistency of their measurements. (Jaeger, 1991, p. x, emphasis added)

However, while generalizability analyses are becoming more commonly used in educational research, this sophisticated technique has not generally been applied to affective measures.

The benefits of generalizability theory have been highlighted elsewhere (e.g., Shavelson & Webb, 1991; Thompson, 1991; Webb, Rowley & Shavelson, 1988), but several of these benefits have not been sufficiently recognized in contemporary analytic practice. A brief review of salient aspects of generalizability theory may be helpful.

Generalizability theory, which subsumes classical test score theory as a special case, also extends classical test theory by recognizing and estimating the magnitude of the *multiple* sources of error (Brennan, 1983; Eason, 1991; Shavelson & Webb, 1991). Both sources of error variance and interactions among these sources can be considered *simultaneously* in a single generalizability analysis. Classical theory admits consideration of only one type of measurement error at a time, and does not consider the possible, completely independent or separate interaction effects of the sources of measurement error variance.

Simultaneous consideration of multiple sources of error variance and the interactions of these error sources is critical, as Thompson (1991) noted:

> I believe that most measurement classicists *unconsciously* presume [*both*] that their error variance sources (a) substantially overlap each other and (b) do not interact to create additional new error variance. Thus, a practitioner may do classical internal consistency, test-retest, and equivalent forms reliability analyses, and may find in all three that measurement error comprises 10% of score variance. Too many classicists would tend to assume that these 10 percents are the same and also tend to not realize that in addition to being unique and cumulative, the sources may also interact to define disastrously large interaction sources of measurement error not considered in classical theory. The effects of these assumptions are all the more pernicious because of their unconscious character. (Thompson, 1991, pp. 1071-1072)

Since the goal of research is usually to generalize over items, occasions, test forms, administrations, etc., generalizability theory honors the reality to which we wish to generalize. As Thompson (1991) noted,

> too few researchers recognize that in all analyses we inherently invoke both a presumptive model of reality and an analytic model. When the two don't match, the analysis doesn't help us understand the reality we believe exists. If we virtually always

> want to generalize over time and over items or tests, then a classical theory approach that never simultaneously considers these two time and item sampling influences, and completely ignores the interactions of these influences, will be quite simply unworkable! (Thompson, 1991, p. 1072)

Thus, generalizability theory could shed a new light on the measurement of affective constructs, such as depression, and provide information with which current measurement techniques could be improved.

Generalizability analyses can also be used conduct so-called "D(esign)-studies" to address important "what if" questions about variation in measurement design (Thompson & Melancon, 1987). Sources of error can be pinpointed and protocol modifications specified that will result in the desired level of generalizability can be elaborated.

Lastly, decisions made in the context of cutoff scores (absolute decisions), as against decisions only considering stability in a relative standing can be considered. Classical test theory does not admit a distinction between reliability involving *absolute* decisions made in the context of cutoff scores (e.g., intervention decisions invoking a cutoff score for severe depression) as against reliability involving decisions only considering stability in *relative* standing or rankings. This distinction can be important, particularly when decisions regarding intervention are involved. In a clinical case, the relative

standing of a child compared to the child's peers is considerably less important than the standing of the child compared to a clinically valid criterion. In generalizability studies the coefficients that address reliability in the context of *relative* decisions, i.e., decisions only concerned with the stability of score rankings, are called *generalizability* coefficients. The coefficients that address reliability in the context of *absolute* decisions, i.e., decisions invoking cutoff score criteria, are called *phi* coefficients.

Realistically, the goal of assessing a child's level of depressive symptomatology is to obtain data that can be generalized simultaneously over items, occasions (time), tests, and administrations. Generalizability theory focuses on the simultaneous influence of multiple sources of measurement error variance, and therefore more closely fits the interests of researchers and clinicians. Still, regardless of the strengths of generalizability theory, the theory has not yet been widely applied to affective measures or, specifically, to measures of depression. Therefore, in the present study we investigated the psychometric properties of a commonly used measure of childhood depression using both classical and generalizability test theories. A better model of phenomena involving childhood depression may result from such investigations, and clinical interventions are improved as we gain insights into our measures.

Method

Data from 164 children from small Texas communities provided

the basis for our analyses. The sample consisted of a few more females (51%) than males, and the children ranged in age from 11-16 years old with a mean age of 12.6. At the first point of measurement the subjects were from grades five (15.3%), six (14.7%), and seven (70.1%). The ethnic composition of the sample was: African American (43.6%), White (38.8%), and Hispanic (16.4%). Subjects were tested on a second occasion after a 28-week interval.

The Children's Depression Inventory (CDI; Kovacs, 1981) was selected for the present study based on its frequency of use and more thoroughly explored psychometric properties (e.g., Finch & Saylor, 1984; Reynolds, Anderson, & Bartell, 1985; Romano & Nelson, 1988). The CDI is the most commonly used self-report measures for both clinical and research purposes. It is a 27-item, symptom-oriented scale suitable for use with children aged 6 to 17 years old. Item responses are scored 0, 1, or 2, with a higher score indicating a more depressed response. A total score greater then 11 has been taken to suggest at least mild depression (Kaslow, Rehm, & Siegel, 1984), while a score greater than 19 suggests severe depression (Smucker, Craighead, Craighead, & Green, 1986).

Previous reliability studies employing classical test theory have reported alpha coefficients in the .70's and.80's. For example, Smucker, Craighead, Craighead, and Green (1986) calculated alphas of .84 and .87 for male and female students in grades three through six, respectively, of .83 and .85 for male and female students in grades seven through nine, respectively, and of .89 for both genders for another sample of students in grades six through

eight. Kovacs (1981) reported a coefficient alpha of .86 in a sample of children and adolescents in a variety of diagnostic categories and alphas of .71 and .87 in samples of pediatric medical outpatients and public school students, respectively. Weiss (1990) reported alphas of .86 for children and .88 for adolescents from samples seeking treatment at 19 mental health facilities in 19 states. Thus, Kazdin (1990) characterized the CDI as having reasonable internal consistency.

With respect to stability of scores over time, Kovacs (1986) reports a correlation coefficient of .82 over a four-week period with a small sample of diabetic children, and of .84 over a nine-week period for a sample of public school children. Kaslow, Rehm, and Siegel (1984) report a test-retest correlation of .83 for a sample of elementary school children over a three-week period. Smucker, Craighead, Craighead and Green (1986) reported test-retest correlations ranging from .74 to .77 for fifth graders after three weeks and, over the span of one year, correlations from .41 to .69 for seventh and eighth graders. Over a 16-week period, Weiss (1990) found test-retest correlations of .54 for children and .56 for adolescents. Clearly, data from the CDI have been fairly thoroughly investigated using classical test theory.

Results

Table 1 presents the variance components from the generalizability analysis. The two major sources of variance were the Persons x Items interaction (17.3%) and the Persons x Time x Items interaction (58.0%). The Persons x Items interaction

suggests that the CDI items tended to be interpreted differently by different persons. The Persons x Time x Items interaction term was the unique combination of person, time, and items as well as unmeasured sources of error variation. Time with $\underline{k}$=2 contributed negligibly (0.2%) to the score variance. Similarly, the two-way interaction effects involving time contributed relatively little variance to scores.

Insert Table 1 about here

Table 2 presents the results from classical test score and generalizability analyses. The Table 2 results utilize the variance components reported in Table 1 to derive the coefficients for the different measurement protocols. Both generalizability coefficients, associated with relative decisions, and phi coefficients, associated with absolute decisions (e.g., cutoff scores), are presented.

Insert Table 2 about here

Discussion

The Table 2 results reflect the fact that classical and generalizability theories can yield different estimates, even for the same data. For example, internal consistency reliability for CDI scores ranged from .86 to .88. These results are consistent with those previously reported in the literature. However, the

generalizability of these data was considerably lower, .63. This suggests that the various sources of error variance are not independent and do interact to markedly decrease the dependability of the scores. It is also noteworthy that the coefficients for both relative and absolute decisions tended to be so comparable, but that both values (.63 and .61) were somewhat small. These results suggest that the CDI must be used with some caution by clinicians and practitioners regardless of whether CDI scores are used for making absolute decisions (e.g., decisions to intervene using score cutoffs) as against relative decisions only focusing on the stability of rankings.

As noted in Table 2, one interesting finding of the present study is that the generalizability of CDI scores increases noticeably over testing occasions, going from .63 with a single testing occasion to .81 with three testings. Conversely, using a single testing occasion, even when the number of items is as high as 108, the generalizability remains relatively low, i.e., .69. In each of the measurement protocols considered, the dependability of the depression scores increases appreciably with an increase in testing occasions.

These results have several possible implications for clinical practice and early intervention with children at risk for depression. Most assessments of depression in children focus on collecting data at a single point in time and estimating the stability of the results. However, these present results suggest that evaluation of depression in children might better invoke

planned, multiple assessments over time. The dependability of the depression scores, given testing with 27 items at a single occasion of measurement, is relatively low. However, testing over multiple occasions will more likely yield data that can be generalized across time, items, and situations.

An example may be illustrative. If, for instance, a child scores in the severely depressed range on the CDI, the clinician will probably suggest an intervention. This decision may be based less on the stability of the symptoms than on the emotional "cost" of the symptoms. As has been documented in the literature, depressive episodes tend to be self limiting by nature and many times will remit without intervention (Beck, 1967; Elkin et al., 1989; Robins & Guze, 1972). Thus, the clinician's decision to intervene will likely focus less on the genesis of symptomology and more on remediation of the psychological distress felt by the child and on prevention of possible negative outcomes associated with this level of depressive symptomatology, most notably suicide. Additionally, intervention may shorten the duration of the depressive episode and reduce the likelihood of recurrent episodes.

If a second child scores in the mildly depressed range, he/she will most likely not be referred for treatment. The depressive symptoms will not, as yet, be considered a cause for immediate concern since they are not necessarily stable. However, if in a planned second or third testing over time this same child continues to report mild (or moderate) depressive symptomatology, it is more likely that these symptoms are stable and this child would then

also be a candidate for intervention.

The present results suggest that screening for early symptoms of depression with repeated testings may allow intervention at an early stage and preventing the development of more serious depressive disorders. With each testing, intervention strategies can then be twofold. First, those children who have extremely high scores at any occasion and are therefore likely experiencing moderate to severe psychological distress can be the target of immediate intervention. Secondly, children whose scores over time reflect consistent mild depressive symptomatology can be targeted for intervention geared primarily at preventing the intensification of depressive symptoms as well as addressing the existing symptoms.

References

Beck, A. T. (1967). Depression: Clinical, experimental, and theoretical aspects. New York: Hoeber.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Carlson, G. & Cantwell, D. (1980). A survey of depressive symptoms, syndrome, and disorder in a child psychiatric population. Journal of Child Psychiatry and Psychology, 21, 19-25.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnum, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples (Vol. 1, pp. 83-98). In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments. Greenwich, CT: JAI Press.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., et al. (1989). National Institute of Mental Health treatment of depression collaborative research program: General effectiveness of treatments. Archives of General Psychiatry, 46, 971-982.

Finch, A. J., Jr., & Saylor, C. F. (1984). An overview of child depression. In W. Burns & J. V. Lavigne (Eds.). Progress in pediatric psychology. Vol I (pp. 201-239). New York: Grune

& Stratton.

Jaeger, R. (1991). Foreword. In R.J. Shavelson & N.M. Webb, Generalizability theory: A primer (pp. ix-x). Newbury Park: SAGE Publications.

Kaslow, N.J., Rehm, L.P., & Siegel, A.W. (1984). Social-cognitive and cognitive correlates of depression in children. Journal of Abnormal Child Psychology, 12, 605-620.

Kazdin, A. E. (1990). Assessment of childhood depression. In A. M. La Greca (Ed.), Through the eyes of the child: Obtaining self-reports from children and adolescence (pp. 189-233). Needham Heights, MA: Allyn and Bacon.

Kovacs, M. (1981). Rating scales to assess depression in school aged children. Acta Paedopsychiatrica, 46, 305-315.

Kovacs, M. (1986). The Children's Depression Inventory. In D. J. Keyser & R. C. Sweetland (Eds.), Test Critiques (Vol. 5, pp. 65-72). Kansas City, Missouri: Test Corporation of America.

Reynolds, W. M., Anderson, G., & Bartell, N. (1985). Measuring depression in children: A multi-method assessment investigation. Journal of Abnormal Child Psychology, 13, 513-526.

Robins, E., & Guze, S. B. (1972). Classification of affective disorders: The primary-secondary, the endogenous-reactive, and the neurotic-psychotic concepts. In T. A. Williams, M. M. Katz, & J. A. Shields (Eds.), Recent advances in the psychobiology of the depressive illnesses (pp 283-293). Washington, DC: U.S. Government Printing Office.

Romano, B. A., & Nelson, R. O. (1988). Discriminant and concurrent validity of measures of children's depression. Journal of Clinical Child Psychology, 17, 255-259.

Shavelson, R. J. & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park: SAGE Publications.

Smucker, M. R., Craighead, W. E., Craighead, L. W., & Green, B. J. (1986). Normative and reliability data for the Children's Depression Inventory. Journal of Abnormal Child Psychiatry, 14, 25-39.

Thompson, B. (1991). Review of Generalizability theory: A primer by R.J. Shavelson & N.W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

Thompson, B. & Melancon, J. G. (1987). Measurement characteristics of the Group Embedded Figures Test. Education and Psychological Measurement, 47, 765-772.

Webb, N.M., Rowley, G.L., and Shavelson, R.J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Weiss, B. (1990). Developmental differences in the factor structure of the Children's Depression Inventory. Unpublished manuscript.

Worchel, F., Nolan, B., & Willson, V., (1987). New perspectives on childhood depression. Journal of School Psychology, 25, 411-414.

Table 1
Variance Sources and Their Proportional Contributions

Variance Source	Variance	% Variance
Systematic		
Persons	0.04318	13.4%
Measurement Error		
Time	0.00060	0.2%
Items	0.01863	5.8%
Persons x Time	0.01665	5.2%
Persons x Items	0.05582	17.3%
Time x Items	0.00030	0.1%
Persons x Time x Items	0.18703	58.0%

Table 2
Classical and Generalizability Coefficients

Facets	Classical r	Generalizability[a]	Phi[b]
Time (k=1) Items (v=27)	.86 & .88[c]	.63	.62
Time (k=2) Items (v=27)	.66	.75	.74
Time (k=3) Items (v=27)		.81	.80
Time (k=1) Items (v=54)		.67	.66
Time (k=2) Items (v=54)		.79	.78
Time (k=3) Items (v=54)		.84	.84
Time (k=1) Items (v=81)		.68	.68
Time (k=2) Items (v=81)		.81	.80
Time (k=3) Items (v=81)		.86	.85
Time (k=1) Items (v=108)		.69	.68
Time (k=2) Items (v=108)		.81	.81
Time (k=3) Items (v=108)		.87	.86
Time (k=1) Items (v=13)		.55	.53
Time (k=2) Items (v=13)		.69	.67
Time (k=3) Items (v=13)		.75	.73

[a]The generalizability to the score universe (reliability) for relative decisions, i.e., decisions considering only rank orderings.

[b]The generalizability to the score universe for absolute decisions, i.e., decisions invoking cutoff scores.

[c]Alpha coefficients for times one and two, respectively.

Appendix A

SUMMARY OF SELECTED D STUDY RESULTS

D STUDY DESIGN NO.	MEASUREMENT OBJECT $P INF.	FACETS T INF.	FACETS I INF.	VARIANCES UNIVERSE SCORE	VARIANCES LOWER CASE DELTA	VARIANCES UPPER CASE DELTA	VARIANCES GEN. COEF.	VARIANCES PHI
001		1	27	.04342	.02584	.02722	.62686	.61469
002		2	27	.04342	.01396	.01500	.75665	.74325
003		3	27	.04342	.01000	.01093	.81273	.79896
004		1	54	.04342	.02136	.02237	.67029	.65992
005		1	81	.04342	.01986	.02076	.68613	.67652
006		2	54	.04342	.01120	.01188	.79495	.78513
007		2	81	.04342	.01028	.01084	.80859	.80016
008		2	108	.04342	.00982	.01032	.81559	.80789
009		2	13	.04342	.01992	.02171	.68551	.66667
010		3	13	.04342	.01472	.01640	.74678	.72586

Note. The universe score variance is the variance in scores attributable to the "object of measurement", here people, and is considered systematic and not measurement error variance. Lower case delta is the variance associated with the pooling of all sources of error variance that would affect relative decisions; upper case delta is the variance associated with the pooling of all sources of error variance that would affect absolute decisions. The generalizability coefficient is the ratio of systematic variance to total variance, i.e., total variance is the systematic variance plus the error variances that impact *relative* decisions; thus the G coefficient for a measurement protocol involving 3 occasions of measurement with 13 items is equal to: .043 / (.043 + .015) = .043 / .058 = .747. The phi coefficient is also the ratio of systematic variance to total variance, but here total variance is the systematic variance plus the error variances that impact *absolute* decisions; thus the phi coefficient for a measurement protocol involving 3 occasions of measurement with 13 items is equal to: .043 / (.043 + .016) = .043 / .060 = .726. Eason (1991) provides more details.