DOCUMENT RESUME

ED 430 049 TM 029 793

AUTHOR	Fan, Xitao; Chen, Michael
TITLE	When Inter-Rater Reliability Is Obtained from Only Part of a Sample.
PUB DATE	1999-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); Generalizability Theory; *Interrater Reliability; *Sample Size; Sampling; Scores; *Scoring

ABSTRACT

It is erroneous to extend or generalize the inter-rater reliability coefficient estimated from only a (small) proportion of the sample to the rest of the sample data where only one rater is used for scoring, although such generalization is often made implicitly in practice. It is shown that if inter-rater reliability estimate from part of a sample is available, the score reliability for the rest of the sample data rated by only one rater can be estimated both within the classical reliability theory framework, and within the framework of generalizability theory. As intuitively expected, score reliability for the data for which only one rater is used for scoring is always lower than the score reliability for the portion of sample data for which two raters are used. A sample of published studies is provided from difference disciplines that gives inter-rater reliability coefficients obtained from a small proportion of a sample. For this sample of published studies, by applying the method discussed in this paper, the estimated score reliability is given for the data rated by only one rater. (Contains 1 table and 20 references.) (Author/SLD)

RATER.V1

When Inter-Rater Reliability Is Obtained from Only Part of a Sample

Xitao Fan

Utah State University

Michael Chen

University of Mississippi

Running Head: Inter-Rater Reliability

Please send correspondence about this paper to:

Xitao Fan
Department of Psychology
Utah State University
Logan, UT 84322-2810

Phone: (435)797-1451
Fax: (435)797-1448
E-Mail: fafan@cc.usu.edu

Paper presented at the 1999 Annual Meeting of the American Educational Research Association, April 19-23, Montreal, Canada (Session # 36.14).

Abstract

It is erroneous to extend or generalize the inter-rater reliability coefficient estimated from only a (small) proportion of the sample to the rest of the sample data where only one rater is used for scoring, although such generalization is often made implicitly in practice. It is shown that if inter-rater reliability estimate from part of a sample is available, the score reliability for the rest of the sample data rated by only one rater can be estimated both within the classical reliability theory framework, and within the framework of generalizability theory. As intuitively expected, score reliability for the data for which only one rater is used for scoring is always lower than the score reliability for the portion of sample data for which two raters are used. We provide a sample of published studies in different disciplines that provided inter-rater reliability coefficients obtained from a small proportion of a sample. For this sample of published studies, by applying the method discussed in this paper, we provided the estimated score reliability for the data rated by only one rater.

In social and behavioral science in general, and in educational and psychological research in particular, there are often situations in which the scoring process is not objective, i.e., the same behaviors will result in different scores if the behaviors are scored by different raters or observers. Within the framework of classical reliability theory, it is usually necessary in these situations to assess the inter-rater reliability of the scores. Inter-rater reliability coefficient provides an quantitative estimate for the amount of measurement error caused by the scoring inconsistency of the raters. For example, in a situation where two raters have independently rated a sample of subjects on some behavior of interest (e.g., performance in an oral exam), the inter-rater reliability coefficient for the data can be obtained by calculating the correlation coefficient between the ratings of the two raters. Let's assume that the result is .80. This inter-rater reliability coefficient can be interpreted to mean that 80%[1] of the observed score variance is due to true score variance (true differences among the subjects on the behavior of interest), and 20% of the observed score variance is error variance due to scoring inconsistency of the two raters (Anastasi & Urbina, 1991, Chapter 4) .

Many practitioners in educational/psychological research do not realize, however, that the interpretation provided above for an inter-rater reliability coefficient is only valid when the average (or the total) of the two scores from the two raters is used to represent each subject's score. In other words, if we use the average (or the total) of the two ratings provided by the two raters for each subject to represent the subject's score, 20% of the variance in these scores is error variance due to rater inconsistency, the remaining 80% of the variance is true score variance, and

[1] It is noted here that reliability coefficient theoretically reflects the ratio between true score variance and observed score variance; as such, reliability coefficient, which takes the form of a statistical correlation coefficient, should not be squared again. Interested readers may see Crocker and Algina (1986, Chapter 6) for details.

the score reliability is 0.80. But if we decide to use only one rater's rating to represent each subject's score, the score reliability is no longer .80, and it will certainly be lower.

The situation described above is not any different from, say, an internal consistency reliability coefficient of .80 estimated for a 40-item test. This reliability estimate is only relevant if we actually use the mean (or the total) of the 40-item test to represent each examinee's score. If we decide that we will only use 10 items (a random sample from the 40 items) as a shortened version, rather than all the 40 items, we no longer can say that the reliability estimate for our shortened version is 0.80. What, then, is the estimated reliability of our shortened version of 10 item-test? Although we may not know it at this time, we are reasonably sure that it will be lower than 0.80.

To estimate inter-rater reliability can often be labor-intensive, and consequently, may be too expensive for a research project. Because of a variety of practical constraints in research (e.g., lack of time, money, or other resources), it is a common practice that some researchers obtain the inter-rater reliability estimate from only a small proportion of their samples. For example, it is not unusual to encounter research studies in which only 10% to 15% of the total sample or observation sessions were rated by two independent raters, and this sub-sample is used to derive the inter-rater reliability estimate (e.g., Bornstein & Tamis-LeMonda, 1990; Carter & Moran, 1991). The rest of the sample, however, is only rated by one rater, rather than two. In this situation, although the score reliability (or amount of error variance) is known for the part of the sample for which two ratings are available, the questions may be asked, "What is the score reliability for the rest of the sample data for which only one rating is available for each observation?".

By itself, the use of a portion of a sample to derive inter-rater reliability coefficient within the framework of classical reliability theory does not cause any methodological problems. But in practice, the interpretation of such a reliability estimate thus obtained is often problematic. The major problem in this situation can be phrased into this question: should this estimate be interpreted as the score reliability for the entire sample, or should the reliability interpretation be limited only to the small portion of the sample from which the inter-rater reliability assessment has actually been conducted? In research practice, the inter-rater reliability estimate obtained from a portion of a sample is usually generalized, although often implicitly, to the entire sample, as if the entire sample has been rated by two raters or observers.

Methodologically, however, such generalization is incorrect, because the obtained rater reliability estimate is for the mean (or the total) score of the two ratings from the two raters. Statistically, such average (or total) scores across two raters tend to be more stable (i.e., more reliable) than scores provided by only one rater. Consequently, the reliability for the rest of the sample data that have been rated by only one rater would be lower, and the inter-rater reliability coefficient derived from only a part of a sample cannot be generalized to represent the score reliability for the rest of the sample data that have been rated by one rater.

If the generalization of inter-rater reliability estimate from a portion of a sample to the entire sample is inappropriate, then how can the reliability for the data of the rest of the sample, for which only one rater is used, be estimated? This problem can be solved both through the classical reliability theory, and through the more versatile generalizability theory. The goal of this paper, therefore, is to illustrate how score reliability estimate can be obtained for the portion of the sample for which only one rater is used instead of two, based on the portion of the sample for which ratings from two raters are available. A brief review of the classical reliability theory and generalizability theory is provided here to lay the groundwork. More detailed discussion of both classical reliability theory and the generalizability theory are provided elsewhere (e.g., Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1971; Eason, 1989; Goodwin, Sands, & Kozleski, 1991; Margery, 1996; Shavelson & Webb, 1991; Thompson & Crowley, 1994).

Classical Reliability Theory

The major question that classical reliability theory poses is how accurately an observed score reflects its corresponding true score. For this purpose, except the true individual differences (true score variance), all other sources of score variation (e.g., items, occasions, raters) are treated as measurement error sources. These different measurement error sources, however, cannot be separated simultaneously. Usually, only one source of measurement error or one undifferentiated error term can be determined at any given time. This undifferentiated error term is one of two parts of the score variance that can be partitioned, the other being the systematic or true variance (true individual differences).

Thus, the observed score can be decomposed into only two parts: true score and error: $X_p = T_p + E_p$, where X is the observed score and the subscripts p refers to persons. The true score, T_p, gives rise to the true score variance (σ_T^2), the observed score, X_p, gives rise to observed score variance (σ_X^2), and the error, E_p, gives rise to error variance (σ_e^2). Because true score and error are independent of each other (i.e., no covariance between the two, we have the relationship of $\sigma_X^2 = \sigma_T^2 + \sigma_e^2$. The theoretical reliability is the ratio of true score variance to observed score variance:

$$r_{XX'} = \sigma_T^2/\sigma_X^2 = \sigma_T^2/(\sigma_T^2 + \sigma_E^2)$$

In practice, because true score variance is never known, theoretical reliability is usually estimated as a correlation coefficient. For example, in a situation where two raters rated the same sample of subjects on some behavior of interest, the reliability for the average ratings across the two raters is estimated by calculating the correlation coefficient between the ratings of the two raters. But for a situation where a sample of subjects were rated by two raters on two different occasions, classical reliability theory does not provide any mechanism for simultaneously estimating both the measurement error due to inconsistent scoring by two raters, and the measurement error due to inconsistent scores across two times.

Classical reliability theory provides some limited flexibility in estimating score reliability for different measurement protocols. For example, if we estimated that the reliability estimate for a 40-item test is 0.80, what would be the approximate score reliability if we decided to use only ten items rather than all 40 items, assuming that the ten items were a random sample of the forty items? For this purpose, the generalized Spearman-Brown formula (Traub, 1994, Chapter 7) can be used to obtain an estimate of score reliability for our planned 10-item test. Generalized Spearman-Brown formula takes the form:

$$\rho_X^2 = \frac{k\rho_Y^2}{1 + (k-1)\rho_Y^2}$$

where, ρ^2_X is the estimated score reliability for the new test, while ρ^2_Y is the computed reliability estimate of the original test, and k is the factor of test length change. If the planned new test is contains twice as many items as the original test, k=2; if the planned new test contains half the items as the original test, k=0.5. In our case, the planned new test is one fourth of the length of the original test, so k=.25, and the estimated score reliability for the planned new test with only 10 items will be:

$$\rho_X^2 = \frac{0.25 \times 0.80}{1 + (0.25 - 1) \times 0.80} = 0.50$$

Generalizability Theory

The major question that generalizability theory poses is the degree of accuracy when the researcher generalizes the observed data to a well-defined measurement universe (e.g., across raters, occasions, items). To this end, generalizability theory (1) permits the simultaneous estimation of all relevant measurement error sources (G study), and (2) allows the researcher to estimate the score reliability under different measurement conditions (D study), such as varying the number of items, the number of raters, and/or the number of occasions used in the measurement process.

The simultaneous estimation of multiple sources of error in generalizability theory is achieved through the decomposition of the observed score variance into multiple sources through the use of analysis of variance (ANOVA) model. As discussed in Shavelson and Webb (1991), in a situation where a sample of subjects (p) were rated by two raters (r) on two different occasions (o), the observed score of a person (X_{pro}) can be decomposed into multiple components that include all the main effects (assuming that persons [p] are the object of measurement, and raters [r] and occasions [o] are the two facets of concern, i.e., two potential measurement error sources), as well as their interactions with each other, plus the residual that contains the three-way interaction term p*r*o:

$$\begin{aligned}
X_{pro} = {} & \mu + && \text{[grand mean]} \\
& (\mu_p - \mu) + && \text{[person effect]} \\
& (\mu_r - \mu) + && \text{[rater effect]} \\
& (\mu_o - \mu) + && \text{[occasion effect]}
\end{aligned}$$

$(\mu_{pr} - \mu_p - \mu_r + \mu) +$	[person-rater interaction effect]
$(\mu_{po} - \mu_p - \mu_o + \mu) +$	[person-occasion interaction effect]
$(\mu_{ro} - \mu_r - \mu_o + \mu) +$	[rater-occasion interaction effect]
Residual (p*r*o, e)	[three-way interaction plus residual]

From this model, the score variance can be decomposed into multiple variance components that represent all the effects (both main and interaction effects):

$$\sigma^2(X_{pro}) = \sigma_p^2 + \sigma_r^2 + \sigma_o^2 + \sigma_{pr}^2 + \sigma_{po}^2 + \sigma_{ro}^2 + \sigma_{pro,e}^2$$

where, σ_p^2, σ_r^2 and σ_o^2 are the variance components for persons, raters, and occasions respectively. σ_{pr}^2, σ_{po}^2, and σ_{ro}^2 are the variance components for the three two-way interactions, and $\sigma_{pro,e}^2$ is for the three-way interaction term confounded with the residual. The generalizability coefficient, which is the conceptual equivalent of the classical reliability coefficient, is the ratio of the variance component of the object of measurement (in most measurement situations, the object of measurement is persons) to the sum of variance component of the object of measurement and the error variance component:

$$\rho^2 = \sigma_p^2/(\sigma_p^2 + \sigma_e^2).$$

Depending on the type of decisions (relative versus absolute decisions) one is interested in making, and on the design of the D study, the error variance component σ_e^2 may consist of different components.

Once the relevant variance components have been estimated through the G study, D study can be conducted either to determine the optimal measurement protocol, or to estimate score reliability under some different measurement conditions (Brennan, 1992; Shavelson & Webb, 1991). In this regard, generalizability theory provides full flexibility (compared to the limited flexibility classical reliability theory offers in this regard) for estimating score reliability of a planned measurement protocol that may be different from the G study design on multiple dimensions (e.g., simultaneously changing both the number of raters and the number of occasions). The flexible and versatile generalizability theory model, in fact, subsumes all other reliability estimates within classical reliability theory as special cases (Eason, 1989).

Methods and Procedures

At the beginning of this paper, we asked the question: if inter-rater reliability estimate is obtained from a portion of a sample, what is the score reliability for the rest of the sample data for which only a single rating from one rater is available for each subject (observation)? Because only one source of potential measurement error exists in this situation (i.e., rater inconsistency), the answer to this question can be obtained either through classical reliability theory, or through generalizability theory. The following sections provide details of solutions to the question.

Solution from Classical Reliability Theory

Researchers generally understand that the generalized Spearman-Brown formula is used for estimating the impact of test length change (i.e., increase or reduction of the number of items on a test) on score reliability. Many of them, however, do not realize that the generalized Spearman-Brown formula is equally applicable in situations that involve the change in the number of raters (Crocker & Algina, 1986, p. 167).

In the situation where inter-rater reliability coefficient has been obtained from part of a sample, and we are interested in estimating the score reliability for the rest of the sample data on which rating from only one rater is available, the generalized Spearman-Brown formula can be used. As discussed before, the generalized Spearman-Brown formula takes the form:

$$\rho_X^2 = \frac{k\rho_Y^2}{1 + (k-1)\rho_Y^2}$$

In our situation, ρ^2_X is the estimated score reliability for the data for which only a single rating from one rater is available for each subject, and ρ^2_Y is the obtained inter-rater reliability coefficient for the part of the sample data for which two ratings from two independent raters are available for each subject. In this case, k=0.5, because there is 50% reduction in the number of raters. Let's assume that for the part of the sample for which two raters rated each subject, the inter-rater reliability coefficient obtained is 0.80. Using the generalized Spearman-Brown formula, the estimated score reliability for the rest of the sample data for which only one rater rated each subject is:

$$\rho_X^2 = \frac{0.5 \times 0.8}{1 + (0.5-1)0.8}$$
$$= 0.67$$

The results here indicate that, if each subject is rated by two raters, and the average (or the total) of the two ratings is used as the score for each subject, 80% of the score variance is attributable to true score variance, and 20% of the score variance is error variance. But for the proportion of the sample data for which only a single rating from one rater is available for each subject, approximately 67% of the score variance is attributable to true score variance (true individual differences), and about 33% of the score variance is error variance due to potential rater inconsistency. In other words, the use of a single rater reduces the score reliability, as we intuitively expect.

Solutions from Generalizability Theory

It is well known that the generalizability coefficient for relative decisions can be estimated from:

$$\rho^2_{rel} = \sigma_p^2 / (\sigma_p^2 + \sigma_{rel}^2),$$

where σ_p^2 is the variance component for the object of measurement (in most applications, person), and σ_{rel}^2 is the error variance for relative decisions. For a one-facet design with rater as the only measurement error source, we have the following (Shavelson & Webb, 1991):

$$\sigma^2_{rel} = \sigma^2_{pr,e} / n_r.$$

where, n_r. represents the number of raters. If n_r=2 (two raters), the generalizability coefficient thus obtained is equivalent to the inter-rater reliability coefficient obtained from classical reliability theory. Thus, for a situation with two raters, and the inter-reliability coefficient of, say, .80, it is possible to solve the equation for the generalizability coefficient, calculate the value of σ^2_{rel} and substitute this value into the new equation for estimating score reliability when a single rating from one rater is available for each observation.

Going back to our earlier example, if a inter-rater reliability of 0.80 is obtained from part of sample data, it means that the generalizability coefficient based on two raters for relative decisions is 0.80 (ρ^2_{rel} = .80). Put this generalizability coefficient into the formula, we have:

$$.80 = \sigma_p^2/(\sigma_p^2 + \sigma_{rel}^2) = \sigma_p^2/(\sigma_p^2 + \sigma^2_{pr,e} / n_r)$$

We do not know the actual values of σ_p^2 and σ_{rel}^2. But because the ratio of object of measurement variance component (σ_p^2) to the sum of object of measurement variance component plus error variance component ($\sigma_p^2 + \sigma_{rel}^2$) must be 80/100, we can say that proportionately, the following relationship must exist:

$$\rho^2_{rel} = 0.80 = .80/(.80 + .20)$$

where $\sigma^2_{pr,e}/2 = .20$, because inter-rater reliability is based on two raters, $n_r = 2$. Solving for $\sigma^2_{pr,e}$ yields a $\sigma^2_{pr,e}$ of .40. Now, using the equation for the relative decision generalizability coefficient with one rater, we have:

$$\rho^2_{rel} = .80 / [.80 + (.40/1)] = .67.$$

This shows what is intuitively expected: single rating from one rater has lower reliability than averaged ratings based on two raters (i.e., for n_r=2, ρ_{rel} = .80). It is noted that the results are the same whether the solutions are obtained through the generalized Spearman-Brown formula in classical reliability theory or through generalizability theory. As a matter of fact, when only one facet (i.e., one source of measurement error) is in question, the results from classical reliability theory and those from generalizability theory are always the same. It is when multiple facets are present (e.g., raters and occasions) that generalizability theory shows its advantage over classical reliability theory.

Some Examples of Published Research Studies

There are many research studies that reported score reliability in the form of inter-rater reliability based on only part of the sample in the study. The score reliability for the rest of the sample data in the study, however, is generally unknown, because only one rater was used for the rest of the sample data. The method presented in the previous sections is applied to a sample of published research studies that produced inter-rater reliability coefficients based on a small proportion of their respective samples, and estimated the score reliability for the rest of their respective samples for which only one rater was used for scoring. The results are presented in Table 1. As indicated in Table 1, the estimated score reliability for scores provided by one rater is considerably lower than that for the average scores based on two raters.

Insert Table 1 about here

Summary and Conclusions

The purpose of this paper is to illustrate that it is erroneous to extend or "generalize" the inter-rater reliability coefficient estimated from only a (small) proportion of the sample with two raters to the larger sample where only one rater is used, although such generalization is often made implicitly in practice. It is shown that if inter-rater reliability estimate from part of a sample is available, this estimate should not be generalized to the data of the rest of the sample for which only one rater is used for scoring, rather than two raters. But the score reliability for the rest of the sample data can be estimated both within the classical reliability theory framework, and within the framework of generalizability theory. As intuitively expected, score reliability for the data for which only one rater is used for scoring is always lower than the score reliability of the small proportion of the sample data for which two raters are used. We provide a sample of published studies in different disciplines that provided inter-rater reliability coefficients obtained from a small proportion of a sample, but implicitly generalized such reliability estimate to the data of the entire sample. By applying the method presented in this paper, we provided the estimated score reliability coefficients for the data rated by only one rater for this sample of published studies.

It should be noted, however, that both classical reliability theory approach and generalizability theory approach can be used in this situation, because only one source of measurement error (one facet) is involved. If multiple measurement error sources are of interest (e.g., both rater and occasion), then the classical reliability theory approach will fall short, and generalizability theory approach is the only viable approach for score reliability estimation. In light of the fact that classical reliability estimates are actually special cases of generalizability theory, it is somewhat surprising how often classical reliability theory is used in favor of generalizability theory, even when the measurement situation warrants the use of the latter over the former. Indeed, some researchers have advocated placing less emphasis on the use of classical reliability theory, and placing more emphasis on the generalizability theory (Margery, 1996; Sun, Valiga, & Gao, 1997; Thompson, 1991; Weiss & Davison, 1981). Appropriate use of generalizability theory, of course, will depend on deeper understanding of its many statistical complexities and more adequate training in its use and applicability.

References

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Bornstein, M. H., & Tamis-LeMonda, C. S. (1990). Activities and interactions of mothers and their firstborn infants in the first six months of life: Covariation, stability, continuity, correspondence, and Prediction. Child Development, 61, 1206-1217.

Bornstein, M. H., Haynes, O. M., O'Reilly, A. W., & Painter, K. M. (1996). Solitary and collaborative pretense play in early childhood: Sources of individual variation in the development of representational competence. Child Development, 67, 2910-2929.

Brennan, R. L. (1992). Generalizability theory. Educational Measurement, 11, 27-34.

Carter, D. E., & Moran, J. J. (1991). Interscorer reliability for the Hand Test administered to children. Perceptual and Motor Skills 72, 759-765.

Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. Fort Worth, TX: Holt, Rinehart and winston, Inc.

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

Dipietro, L. A, Caspersen, C. J., Ostfeld, A. M., & Nadel, E. R. (1993). A survey for assessing physical activity among older adults. Medicine and Science in Sports and Exercise, 25, 628-642.

Eason, S. H. (1989). Why Generalizability Theory yields better results than Classical Test Theory. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. (ERIC Document Reproduction Service No. ED 314-434).

Goodwin, L. D., Sands, D. J., & Kozleski, E. B. (1991). Estimating inter-interviewer reliability for interview schedules used in special education research. Journal of Special Education, 25, 73-89.

Marcus, B. H., Selby, V. C., Niaura, R. S., & Rossi, J. S. (1992). Self-efficacy and the stages of exercise behavior change. Research Quarterly for Exercise and Sports, 63, 60-66.

Margery, A. E. (1996). Influences on and limitations of Classical Test Theory reliability estimates. Paper presented at the Annual Meeting of the Southwest Educational Research Association. ERIC Document Accession No. ED 395-950.

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

Smith, K. E., Landry, S. H., Swank, P. R., Baldwin, C. D., Denson, S. E., & Wildin, S. (1996). The relation of medical risk and maternal stimulation with preterm infants' development of cognitive, language, and daily living skills. Journal of Child Psychology and Psychiatry and Allied Disciplines, 37, 855-864.

Sun, A., Valiga, M. J., & Gao, X. (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. Journal of Experimental Education, 65, 367-379.

Tamis-LeMonda, C., & Bornstein, M. (1990). Language, play, and attention at one year. Infant Behavior and Development, 13, 85-98.

Thompson, B. (1991). Review of Generalizability theory: A primer. by R. J. Shavelson & N. M. Webb. Educational and Psychological Measurement, 51, 1063-1068.

Thompson, B., & Crowley, S. L. (1994). When classical measurement theory is insufficient and Generalizability Theory is essential. Paper presented at the Annual Meeting of the Western Psychological Association. (ERIC Document Reproduction Service No. ED 377-218).

Traub, R. E. (1994). Reliability for the social sciences: Theory and applications. Thousand Oaks, CA: Sage.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. Annual Review of Psychology, 32, 629-658.

Table 1 Reliability Coefficients from Part of a Sample and the Estimated Score Reliability for the Rest of the Sample - Examples of Published Research Studies

Study	Construct Measured	N (% with Two Raters)	Reported Inter-Rater Reliability for Part of a Sample	Estimated Score Reliability for Data with One Rater
Bornstein & Tamis-LaMonda (1990)	Mother-infant attention and vocalizations	28 (25%)	.92	.85
Bornstein, Haynes, O'Reilly, & Painter (1996)	Maternal Play Solicitations	141 (17%)	.78	.64
Carter & Moran (1991)	Affection in children	679 (15%)	.66	.49
Dipietro, Caspersen, Ostfeld, & Nadel (1993)	Physical activity among older individuals (8 measures)	134 (57%)	.54 (median) (.42-.65)	.37 (median)
Marcus, Selby, Niaura, & Rossi (1992)	Self-efficacy and Stages of Exercise Behavior	429 (4.67%)	.90	.82
Smith, Landry, Swank, Baldwin, Denson, & Wildin (1996)	Maternal Attention-Maintaining Directiveness	340 (20%)	.93	.87
Tamis-LeMonda & Bornstein (1990)	Toddler attention	43 (20%)	.87	.77