Reliability in Comparative and Ethnographic Observations: The Example of High Inference Father-Child Interaction Measures

DOUGLAS R. WHITE

School of Social Sciences
University of California, Irvine
Irvine, CA 92697

ABSTRACT: The theory of reliability and reliability estimates, nearly a century old, has rarely been employed in anthropology, both for lack of familiarity and related problems of computation. This theory is reviewed and considered for use with anthropological data. A set of procedures is provided which combine existing methods to solve the practical problems in use of the theory to assess the reliabilities of composite measurement scales combining multiple measures, of individual independent measurements of a single concept, and of individual cases scored on the composite scales. These procedures are also embodied in a computer program, the results of which are explicated. While domain sampling assumptions are the only requirements of assessing reliability of composite scales, strict assumptions and validation procedures are discussed for the assessment of individual variable reliabilities. An illustration using reliability theory is drawn from cross-cultural studies for "high inference" measures from four different studies of father - child interaction. Validity issues are illustrated both in terms of tests for measurement bias and construct validity for the hypothesized relation between the father - child bond and beliefs in high gods.

Reliability theory is equally applicable in comparative and ethnographic case studies. It offers research practices and theoretical understandings that are capable of integrating and mediating discourse between many of the splintered schools of thought in anthropology, and healing some of the rifts between them. Essentialist biases are discussed as one reason why the theory is not more, widely employed.

KEY WORDS: Measurement, reliability, validity, methods, cross-cultural research, father-child, high gods.

INTRODUCTION

While the use of statistical methods of reasoning and validation has been the hallmark of a scientific approach in social studies (Romney 1989: 168, citing Fisher 1948: 2), ethnography and ethnology, including cross-cultural research, have paid scant attention to the statistical theory and measurement of reliability. This may be due in large part to the high "cost" of acquiring ethnographic and comparative data, and to the assumption that high quality data is best obtained by careful observation and coding without pretense of "replication." But even texts of research methods in

Journal of Quantitative Anthropology 2: 109-150, 1990.

anthropology, while they contain important discussion of reliability and validity (Pelto and Pelto 1978; Bernard 1988: 48-61), often contain no hint of how to do reliability analysis. Assumptions about the "impossibility" of replication may run deeper still.

The logic which justifies the predominant absence of attempts to attain replication in field studies (with notable exceptions), asserts that observations made by anthropologists in the field cannot be replicated since the situations they observe are constantly changing, and no two observers will observe the same situation, much less interpret such situations in identical fashion. This logic, which denies the possibility of replication and reliability to anthropology and is often a version of the relativist argument (Spiro 1986), is fallacious. It thrives on ignorance of the nature of replication, of the theory of reliability which is central to statistical inference and measurement, and of the evolutionary foundations of knowledge (Campbell 1970; Toulmin 1981). It privileges the individual observer and a particularistic and subjectivist conception of knowledge at the expense of convergent or intersubjective knowledge. In contrast, the theory of reliability is based on the direct comparison of multiple measures or observers, employing scientific skepticism to try to subject established results to more strenuous replications. Reliability concerns what is repeatable in measurement, such as different measures of an attribute, or different repetitions of the same measure of an attribute. It is a generic concept, referring to the accuracy (equivalence and stability) of measurement within a particular context of replication, including both the population studied and those engaged in doing the study (Lord and Novick 1968: 139).

This study explicates the measurement of reliability in multiple observations or assessments, following the lead of Romney (1989), in hopes that methods for reliability assessment will be easily assimilated in ethnographic and comparative research. Each of the basic types of reliability problems in anthropology can be seen as a special case of measuring reliability given sampling of a conceptual domain by multiple measurement. This includes the problem of multiple observer reliability where the same situations or materials are independently measured on the same conceptual property, but distinct measures or types of measures 'are independently derived by each observer or investigator. In comparative studies, for example, reliability problems might include different coders independently coding the same conceptual domain, such as severity of alcoholism, or different variables rated by one coder that tap the domain. Another basic type of reliability problem is the case of multi-observer or multi-coder reliability where the same situations or materials are independently measured on the same conceptual property, using shared definitions of the variables (with inter-coder reliability in cross-cultural research as a special case, e.g., Bradley 1989). The measure of informant reliability

(Weller and Romney 1988; also see Kirk and Miller 1986, in the context of qualitative data), is a closely related problem, involving a shift in perspective from "variables" to "informants" and cultural consensus models (Romney, Weller and Batchelder 1986) as the focus of analysis.

The explication begins with the most general concept of reliability in comparable constructs (Tryon 1957) to estimate the reliability of scales formed by averaging multiple measurements, working up to efficient estimates of item reliability by the added assumption of a single common systematic measurement component among the measures. Given this additional assumption of Spearman-Brown single factor theory, weighted scales of items of differing reliabilities can be formed that are more reliable than simple averages, and the reliabilities of these scales can be determined.

RELIABILITY MEASUREMENT

There are five general types of reliability measurements. It is useful to discuss them first in terms of their assumptions:

1. The reliability of an additive scale composed of multiple and not necessarily unidimensional measurements of the same concept (Tryon 1957). This requires only the assumption that the measures sample a conceptual domain of measurement.

2. Similar reliabilities of items in an additive scale of multiple measurements of the same concept. This requires the added assumption formulated by Brown (1910) and Kelley (1924) that each measurement constitutes a "parallel test" with equal reliability (thus, equal standard deviations and uniform intercorrelations of items).

3. Potentially different reliabilities of individual items in a scale of multiple measurements of the same concept. This requires an assumption of a "true score" or single underlying component of measurement, as formulated in the one-factor model of Spearman (1910) and Yule (1922). Measurement errors are assumed to be random and independent. The latter assumption is important, since the items sampled must meet the assumption of strict independence in comparable ways, such as a set of measures each by a different coder or observer (e.g., each on the same project, or each on. distinct projects, so that potential joint influences do not form a systematic measurement error source relative to one another), or all measures by the same coder or observer, but each defined independently so that measurement or coding errors on one variable will not affect systematic errors on another.

4. The reliability of a scale composed of multiple measurements of the same concept where items are weighted to make optimal use of their different one-factor reliabilities (designated ORS for optimal reliability

scaling). The reliability estimates for the scale, based on appropriate adjustments for the weighted sums, depend only on Tryon's (1957) more general assumptions of domain sampling (comparable constructs), not on single factor assumptions.

5. The reliability of additive or ORS composite-scores for individual cases. The use of interval-level measurement requires only the assumption that each variable is normally distributed.

While theories of reliability measurement were formulated at the turn of the century (e.g. Spearman 1904a, 1910, 1927; Brown 1910), it took a half-century for a clear formulation (Tryon 1957) of the fact that basic reliability measurement of composite scales did not depend on either the restrictive assumptions of the Brown-Kelley theory of "parallel tests" (employed for example by Nunnally 1967, 1978) or those of the Spearman-Yule theory of true and error factors.

Although questions of reliability are phrased above in terms of the reliability of variables (arriving finally at an assessment of the reliability of composite measures for individual cases), the whole discussion of reliability measures can be reversed so as to begin with questions of the reliability of the individual respondents, for example, to a set of common knowledge questions that have an unknown answer key, and to end with questions of the reliability of composite "best estimates" about the answer key. This is the approach taken to the "cultural consensus" theory of reliability developed by Romney, Weller and Batchelder (1986; see Weller and Romney 1988; Romney and Weller 1989; Batchelder and Romney 1988, 1989). They deal with reliability questions in a probabilistic framework, and use iterative Bayesian estimates of the reliabilities of informants in relation to consensus about an "answer key."

Unlike the cultural consensus mode, the present approach, which assumes interval level measurement of variables, stays within what has come to be called "classical" test theory, the stock in trade of psychometricians. Straightforward measures for these five types of reliability are as follows:

DEFINITION 1

Scale reliability (additive). The reliability *'It* of the sum of distinct measures of a concept is defined as the correlation between the observed summary score *Xt* and a comparable construct *Xt',* one whose test-samples from the same domain vary on the average as much in their variances and inter-correlations as do the test-samples in the observed *Xt* composite (Tryon 1957: 229-231).

Since the average statistical properties of the construct *Xt'* are by definition those of the observed *Xt,* it is unnecessary to actually measure

the comparable construct or second composite scale to calculate r_u. Tryon (1957: 234) shows four mathematically equivalent ways to compute r_u.

"The Variance Form, variously called Alpha, or L_3, or for dichotomous variables the Kuder-Richardson (or K-R) formula 20.

The Part-Whole Form, a special case of which, is called 'Gullikson's formula.' .

The Individual Variance Form, not reported elsewhere to the writer's knowledge [i.e., derived by Tryon 1957: 248].

The Covariance Form, an approximation to which is known as the Spearman-Brown formula."

Table I**,** reprinted from Tryon (1957: 232), with an example of five distinct ratings or test-samples on 10 individuals, illustrates how each of the four measures of scale reliability are computed, with identical results in each case. Table I also gives the Spearman-Brown approximation ("prophesy") formula, known also as Cronbach's alpha.

DEFINITION 2

Homogeneous item reliability (in an additive scale). The reliability r_u of an item or variable in an additive scale is defined as its correlation with the sum of other scale variables sampled from the measurement domain (Nunnally 1967).[1] Note that with homogeneous item reliability, the reliability of a composite scale necessarily improves with added items sampled from the domain of measurement.

This definition of item reliability is useful when all composite scale items have similar reliabilities, but such estimates are unstable for relatively few items of markedly different reliabilities.[2] With few variables, the reliabilities of composite scales with different items deleted will vary to the extent that the reliability of the deleted item is above or below the average.

To estimate item reliabilities of differing magnitude, psychometricians turn to the single factor model. The one factor model assumes that each item consists of a true score component and an independent random error component. A corollary is that there are no systematic measurement components or error factors that correlate across different items. (Note that Definition 1 did note assume a one factor model.)

DEFINITION 3

Heterogeneous item reliability (single factor). The reliability r_u of an item or variable in the single factor model is the ratio of true-score variance to the actual variance of the measure (Nunnally 1967: 181, 198). A standard

TABLE I

Illustrative Score Matrix, and the Reliability Coefficient, r_{tt}, Calculated by Four Alternative Computing Forms.

Score matrix: $N = 10$, $n = 5$

	Test-sample, X_i							Individual Variance Form:
Ind.	X_1	X_2	X_3	X_4	X_5	X_t	V_{0_i}	
1	6	2	1	0	0	9	4.96	$r_{tt} = 1 - \dfrac{\frac{1}{n-1}(n^2\overline{V}_{0_i} + M_t^2 - n\Sigma M_i^2)}{V_t}$
2	8	6	5	2	4	25	4.00	
3	10	12	7	7	7	43	4.24	$= 1 - \dfrac{\frac{1}{4}[25(4.768) + 600.25 - 5(130.47)]}{140.05}$
4	5	11	11	9	8	44	4.96	
5	6	3	0	0	1	10	5.20	$= 1 - 0.120$
6	11	7	9	6	1	34	11.36	
7	7	7	2	5	5	26	3.36	$= 0.880$
8	4	7	4	4	1	20	3.60	
9	6	3	3	2	4	18	1.84	
10	6	5	1	3	1	16	4.16	
ΣX_i	69	63	43	38	32	245	47.68	$\overline{V}_{0_i} = 47.68/10 = 4.768$
Variance Form:							Σ_i	
M	6.90	6.30	4.30	3.80	3.20	24.50	24.50	$r_{tt} = \dfrac{n}{n-1}\left(1 - \dfrac{\Sigma V_i}{V_t}\right)$
M^2	47.61	39.69	18.49	14.44	10.24	600.25	$130.47 = \Sigma M_i^2$	
$V = \sigma^2$	4.29	9.81	12.21	7.96	7.16	140.05	$41.43 = \Sigma V_i$	$= \dfrac{5}{4}\left(1 - \dfrac{41.43}{140.05}\right)$
						$= V_t$	$\overline{V}_i = 8.286$	$= 0.880$

Table I (continued)

Score matrix: $N = 10$, $n = 5$

Test-sample, X_i

Individual Variance Form:

Part-Whole Form:

$\Sigma X_i X_t$	1804	1892	1427	1245	1035	
r_{it}	0.463	0.940	0.903	0.941	0.793	
$\sigma_i r_{it}$	0.959	2.945	3.156	2.653	2.121	$\Sigma\sigma_i r_{it} = 11.834$

$$r_{tt} = \frac{n}{n-1}\left[1 - \frac{\Sigma V_i}{(\Sigma\sigma_i r_{it})^2}\right] = \frac{5}{4}\left(1 - \frac{41.43}{11.834^2}\right) = 0.880$$

Covariance (and its Approx.) Form (Variance-covariance matrix[a]): 1st entry is r_{ij}
2nd entry is $\sigma_i\sigma_j r_{ij} = c_{ij}$ (diagonals are V_i; $\bar{c}_{ij}$ for off diagonals)

Covariance Form:

X_i					
	X_1	1.000 4.29	0.313 2.03	0.377 2.73	0.270 1.58	0.130 0.72
	X_2	0.313 2.03	1.000 9.81	0.778 8.51	0.923 8.16	0.756 6.34
	X_3	0.377 2.73	0.778 8.51	1.000 12.21	0.848 8.36	0.593 5.54
	X_4	0.270 1.58	0.923 8.16	0.848 8.36	1.000 7.96	0.707 5.35
	X_5	0.130 0.72	0.756 6.34	0.593 5.54	0.707 5.34	1.000 7.16

$$r_{tt} = \frac{n\bar{c}_{ij}}{\bar{V}_i + (n-1)\bar{c}_{ij}} = \frac{5(4.931)}{8.286 + 4(4.931)} = 0.880$$

Covariance Approx. (Spearman-Brown) Form:

$$r_{tt} \doteq \frac{n\bar{r}_{ij}}{1 + (n-1)\bar{r}_{ij}} \doteq \frac{5(0.5695)}{1 + 4(0.5695)} = 0.869$$

Source: Robert C. Tryon, 'Reliability and Behavior Domain Validity,' Psychological Bulletin 54: 232. 1957. In the public domain.

expression (Lord and Novick 1968: 216) for this reliability coefficient is:

(1) $r_{ii} = r_{iT} * r_{jT}$

Here, reliability is the square of the correlation r_{iT} between item i and the "true score" T of the measure, estimated by a single factor score. The argument here is that if the same measurement is taken twice (differing only by a linear random component with the same variance), the assumed identical error terms cancel, and the correlation between one test and the other is the ratio of the true variance to the variance of either test. This result is also derived in Spearman's (1904b) one-factor model, in which the correlation r_{iT} between two measures i and j of a common score is the product of their true score correlations:[3]

(2) $r_{ij} = r_{iT} * r_{jT}$

Principal one-factor analysis (e.g., Schuessler 1971: 109-114, following Hotelling 1933) and MINRES one-factor analysis (see different variants in Comrey 1962, or Harman 1967: 187-210)[4] provide iterative least-squares estimates of n true score correlations such as r_{iT} and r_{jT} given $n(n - 1)/2$ correlations such as r_{ij}. This is equivalent to finding factor loadings b_j (= r_{jT}) for each variable.

A less efficient "triads" method (Lord 1980: 39-40) of estimating one-factor reliabilities derives from Spearman's (1904a: 90) Equations (1) and (2):[5]

(3) $r_{ii} = r_{ij} * r_{ik} / r_{jk}$

The one common-factor model (Kim and Mueller 1978: 67) considers each normalized variable Xj to be a weighted sum of common (F) and unique (u) factors such that:

(4) $X_j = b_j F + d_j u_j \qquad (X = b_j F_i + d_j . u_{jj})$

The set B of b_j are the factor loadings of the variables. F is estimated as $\hat{F}$ by minimizing the following sum of squares:

(5) Minimize Sum(i,j) $(X_{ij} - b_{jF} \hat{F})^2$

With heterogeneous item reliability, the reliability of additive scales does not necessarily improve with more items sampled from the domain of measurement.[6] Optimal procedures for factor weighted scales, however, may guarantee that scale reliability will always exceed that of any individual items, and that sampling of additional items will not lower scale reliability.

DEFINITION 4

Optimal reliability scaling (ORS). Once one-factor reliabilities are known,

the optimal weights, *W*, for combining variables into composite scale scores $\hat{F}$ to minimize Equation (5) are computed from a correlation matrix reproduced from factor loadings *(BB')* with ones (an unreduced matrix) rather than communalities on the diagonal. They are constructed by multiplying the inverse of the reproduced correlation matrix by the *B* factor loadings, thus $W = (BB')^{-1}B$. Scale scores (Kim and Mueller 1978: 67) are given by:

(6) Scale Scores = X W = $X(BB')^{-1}B$

Composite scores derived by these weights constitute a least-squares scale that minimizes the sum of squared differences between each normalized variable X*i* and scale scores weighted by the true score correlations *bj* for each variable (Equations (4-5).

The reliability of composite scales formed by optimal one-factor weighting can be estimated from any of the four equations in Table I with appropriate weighting adjustments. In the Variance Form equations, for example, means and variances must be adjusted by the appropriate *W* weights.

Using least-squares factor scales, scale reliability will be at least as good as that of the most reliable item. Even in extreme cases of unequal reliabilities, the scaling weights in Equation (6) guarantee that even variables with extremely low reliabilities will not diminish the reliability of a scale by being averaged in (introducing more rather than less random error) with highly reliable items. For example, Equation (6) guarantees that variables with reliabilities of 1.0 are used exclusive of other items in factor scale construction. Least-squares factor scales also satisfy the rule of thumb (Kim and Mueller 1978: 72) that if the one-factor model fits the empirical data well, and there are factor loadings (true score correlations) above 0.9 (reliabilities under 0.8), it is counter-productive in deriving scale scores to give them the same weight as items with low loadings.

DEFINITION 5

Reliability of average (or weighted) observations on individual cases (reliability-by-case analysis). How to assess the degree of reliability of a composite-score rating for each individual case? For each case, the extent of agreement between different variables is a measure of the reliability of the scoring for that case. Each case *i* has a mean *mi* and standard deviation s; of *k* normalized scores *Xij:*

(7) $s_i = \text{Sum}(x_{ij} - m_i)^2/k)^{1/2}$

The *Si* also have a mean and standard deviation S (indicating disagreements) around their sample mean *M:*

(8) $S = \text{Sum}(s_i - M)^2/k)^{1/2}$

STOPPED EDITING HERE

Each case thus has a unique z-score *Zj* of disagreement variation around the sample mean *D* of standard deviations or disagreements:

(9) $Z_i = (S_i - D)/S$

Z_i is a relative measure of disagreement, case by case, and thus a suitable measure of the reliability of composite scores for individual cases.

RELIABILITY AND VALIDITY ISSUES

Reliability issues are inseparable from validity issues. Four of the ways in which reliability and validity are connected are discussed here: (1) Reliability is closely related to issues of validity in hypothesis testing; (2) Reliability estimates also require validation in terms of measurement assumptions, as in the present instance where interval level measurement is assumed and transformations of variables may be needed to meet appropriate assumptions; (3) Reliability estimates of individual variables require validation of single-factor assumptions and detection of systematic errors where these assumptions may be violated; (4) Standards of reliability provide guidelines for valid inference.

1. *Relation of Reliability to Hypothesis Testing*

The study of reliability is a necessary starting point for any attempt to provide explanations that account for variability. The amount of variance *(r2)* that regression analysis can account for in a variable cannot exceed the reliability of the variable.? The relation between reliability and predictability was explicitly recognized in Yule's (1897, 1899) contribution to statistical theory for testing causal hypotheses (building on Pearson 1896), which integrated controls for potentially confounding third factors (now known as multiple regression analysis), with theories of reliability and prediction around the concept of error variance (Romney 1989: 174-175). Reliability estimates also provide important criteria for replicating tests of hypotheses, on the expectation that high-agreement ratings for valid hypotheses will correlate more accurately than low-agreement ratings.

2. *Appropriate Levels of Measurement and Transformations of Variables*

The present exposition of reliability measurement has assumed interval level measurement. Computation of product-moment correlations typically assumes interval variables that are normally distributed. If this is not the case, there may be problems of statistical validity in reliability assessment. In many applications the level of measurement is ordinal but results are assumed to be robust in spite of the level of measurement. To be conservative

in the level of measurement issue, a simple expedient is to re-express ordinal measures as with interval values that correspond to normality assumptions. Procedures for normalized re-expression are discussed in Tukey (1977: 103).

3. *Validation of Single-factor Assumptions and Detection of Systematic Errors*

While estimation of reliabilities of heterogeneous items is mathematically well formulated given the assumption of no systematic measurement error, systematic measurement errors are potentially serious in practical applications, especially when they are uncorrelated with the valid component of measurement.[8] Also, in building measurement scales from multiple indicators, common systematic measurement errors may cumulate as readily as the valid component of measurement. Fortunately, both single-factor and multi-factor models are capable of detecting systematic measurement errors common to any substantial fraction of variables.

There are three tests of single-factor structure:[9] (1) In the single-factor model, the sum of squared residuals between actual correlations (and communalities on the diagonal) and reproduced correlations (product of factor loadings) is less than $1/N^{1/2}$, where N is the number of variables. (2) With k variables, the first factor should have well over $1/k$ of the variance and the second factor under $1/k$ (Schuessler 1971: 129). (3) In multiple factor analysis the most reliable indicator of single-factor structure is that the variance accounted for by the first factor is at least three times larger than that of the second (Romney 1989: 189; Lord 1980: 21). It is important not to omit any of these tests to insure the validity of estimates of item reliabilities.

4. *Standards of Reliability as Guides to Valid Inferences*

As rule of thumb (Kim and Mueller 1978: 70-72), factor loadings under 0.3 (reliabilities under 0.1, in the present context) are not substantial: over 90% of their variance is error or: unique variance, and they are often within the range of sampling variation around a mean of zero correlation with the factor scale.

Nunnally (1967: 226) gives the following guidelines for standards for composite scale reliability: "In those important settings where important decisions are made with respect to specific test scores, a reliability of 0.90 is the minimum that should be tolerated, and a reliability of 0.95 should be considered the desirable standard." "Even with a reliability of 0.90, the standard error of measurement is almost one-third as large as the standard deviation of test scores" (Nunnally 1967: 226).

However, "For basic research, it can be argued that increasing reliabilities beyond 0.80 is often wasteful. At that level correlations are

attenuated very little by measurement error" (see Equation (2». "In basic research, the concern is with the size of correlations. .. for which purposes a reliability of 0.80 for different measures involved is adequate."

In any case, it is not just individual variables or items that should be accepted or rejected for various purposes based on reliability analysis, but, where possible, the more reliable composite scales from multiple measurements. Probably the minimum reliability of a composite scale that is useful in hypothesis testing is about 0.70 (30% error variance). Below that level, correlations are significantly attenuated by measurement error, and the error is sufficiently high (given the possibilities for systematic as well as random error) that one is no longer sure what is being measured.

COMPUTATIONAL METHODS IN GENERAL AND IN THE SPECIFIC EXAMPLE

This paper illustrates problems of reliability assessment, and related problems of assessing measurement validity, for a single measurement domain in cross-cultural analysis. A computer program is also described that computes reliabilities of scales and individual variables, as in this application. In the application, as in the computer program, one of the four equivalent methods shown in Table I is used to compute the reliability of additive and optimally weighted composite scales. Item reliabilities are estimated by Equation (2) from correlations between measures of tJ'1e same concept using principal factor methods (equivalent to MINRES). These reliabilities in turn estimate expected correlations between measures of the same concept in the one-factor model, via Equation (2). Factor loadings (square roots of reliabilities b_j serve as the basis for determining a weighting of normalized variables that gives an optimal composite scale, one that has the best reliability for the common measurement component of the variables.

The CORR-REL[10] computer program (White 1990) computes all of the reliability measures discussed above for a set of distinct measures of the same concept. The program reads a raw data *file*, gives the option of re-expressing ordinal variables with normalized interval values, accepts a control variable with high and low values if wanted for selecting subsamples, computes correlations and, assuming independent measures of the same concept, estimates:

(1) additive scale reliabilities by one of the equivalent methods in Table I.

(2) item reliabilities from Equation (2) estimated by principal single factor analysis (MINRES) of optimal factor loadings.

(3) expected correlations by Equation (2) as products of the estimated true score correlations.

(4) standard deviation *s of* the difference between expected and original correlations, with *s* < 1/ *NU2* as the first test *of* the one-factor model.

(5) factor scale score weights by optimal least-squares criteria, as in Equation (6).

(6) the proportional variance *v of* the first factor, with *v* < 1/ *k* as the second test *of* the one-factor model (the third test, by conventional factor analytic packages, is that *v!* > 3 *V2* for the proportional variances *v!* and *V2* on the first and second factors).

(7) optimal composite scale (ORS) reliability by one *of* the equivalent methods in Table I, and as approximated by Cronbach's alpha.

(8) correlations between variables and optimal factor scores.

The program also produces two output variables:

(9) optimal factor scores for each case, from Equation (6).

(10) a classification *of* individual cases into high, medium or low for the agreement among multiple indicators, derived from Equations (7-9).

It is not advised, when using these methods for scale construction, to begin with set *of* variables sampling a measurement domain, and then cull out those with low reliabilities. This can lead to spurious results and makes significance tests *of* factor models meaningless. Problems *of* weighting are handled by factor-scaling methods. Items should be culled, however, when they are found to contain serious problems *of* biased measurement correlated with the common valid component *of* measurements, as this will bias other reliability estimates.

FATHER - CHILD INTERACTION: AN EMPIRICAL EXAMPLE OF RELIABILITY ANALYSIS

Four strictly independent measures *of* father - child closeness or "paternal caretaking" in early childhood (after infancy),[11] coded for 186 societies in the Standard Cross-Cultural Sample (Murdock and White 1969), are examined by methods *of* reliability assessment. This sampling domain for assessing multiple measurements is chosen to illustrate problems *of* both reliability and validity. Potential unreliability *of* measures *of* father - child interaction (Johnson and Behrens 1989) makes this domain especially interesting for analysis. Hypotheses tested uncritically with these data may be open to question on the basis *of* measurement reliabilities, but may also be open to improvement in subsequent testing *of* hypotheses.

The father - child interaction codes are a good case *of* "high inference" variables (Whiting 1981) which make the testing *of* hypotheses difficult, since a considerable proportion *of* the variance could be due to random error or a host *of* extraneous considerations besides the main concept. The problem with high-inference codes is that one may not be able to

discern what the coders had in mind in making inferences (see also Bradley 1989) - although in the present instance we have sufficient information to make good guesses about sources of bias.

The four independent ratings[12] of father - child closeness or paternal caretaking are by Barry and Paxson (1971), Barry, Josephson, Lauer, and Marshall (1977), Rohner and Rohner (1982), and Whyte (1978). The first study coded the role of father in early childhood (ages 2-5) on a five-point scale from close to distant. The second coded the gender of parental caretakers in early childhood separately for boys and girls (ages 4-7) on a five-point continuum from males exclusively to females exclusively (I averaged the separate boy and girl scales). The third study coded the importance of fathers for both boys and girls (ages 3-6) on a four-point continuum from "father rarely the caretaker" to "almost always the caretaker," in answer to the question "Who has major responsibility for the routine daily care, supervision, and discipline of the children?" The fourth study coded authority over the care, handling and discipline of children (ages 5 to leaving home) on a four-point continuum from "virtually monopolized by males" (not necessarily father) to "females have more say or virtually monopolized by females."

The coders for the four studies were distinct. Four coders for the first study did two independent ratings followed by "consensus" ratings: Paxson did one set of ratings; Gaver, Paxson and Obrinsky the other. Three coders for the second study (Josephson, Lauer, and Marshall) did separate and later "consensus" ratings for half the societies (96), and single ratings for the remainder (see Barry, Josephson, Lauer, and Marshall, 1976 and 1977: both sets of codes were done at the same time by the same coders). Six coders for the third study (Bell, Carini, Depenbrock, Foster, Lori and Stokes) did two sets of independent ratings, followed by "corrected" consensus codes. For the fourth study, two coders did independent ratings followed by "corrected" separate codes which were then summed for the final codes. Unfortunately, none of these studies published the ratings of the separate coders nor they have made them available. Only Rohner and Rohner report on intercoder reliability *(r* = 0.70 on the uncorrected codes for 38 societies).

Table IIA, showing the correlations between the variables and the number of cases on which each correlation is based, refers to the principal coders of the first two studies as Paxson and Josephson, but refers to the last two by their senior authors, although they are not the principal coders. Variable numbers in column 1 refer to the *World Cultures* database, which is the compendium of data on the standard sample used in this study. Correlations between the variables are given on the upper right in the matrix whose four rows and columns are labeled Paxson, Josephson, "Rohner" and "Whyte" as explained above. The average correlation of 0.26 between the variables gives a simple average composite scale

(Cronbach) reliability of 0.59, which is unacceptably low by Nunnally's standards.

TABLE IIA
Correlations between four father - child closeness measures

Variable Numbers	Principal Coders	Paxson	Josephson	'Rohner'	'Whyte'
54	Paxson	-	r = 0.40	r = 0.41	r = 0.07
369-370	Josephson	N = 142	-	r = 0.34	r = 0.14
991	'Rohner'	N=60	N=65	-	r = 0.20
614	'Whyte'	N=59	N=65	N=27	-
Total Cases Coded		N= 176	N = 168	N=67	N=68 of 93

Average Correlation of 0.26 gives a simple average composite scale (Cronbach) minimal reliability of 0.59

An item-reliability analysis of this correlation matrix and the raw data, in Table lID, shows the reliabilities for the four variables computed from optimal factor loadings (whose products reproduce the original correlation matrix) of 0.34, 0.31, 0.42, and 0.05. Correlations of each variable and optimal scales of other weighted variables are 0.32, 0.30, 0.41, and 0.09, respectively, and optimal scaling weights are 0.34, 0.40, 0.42, and 0.13. Thus, reliability estimates are highest for Paxson's (0.34) and Rohners' variable (0.42), but Josephson's variable also gets high scaling weight (0.40).

The data in Table lIB fail all three tests of single-factor structure. The standard deviation of differences between actual and reproduced correlations is 0.05, which is barely less than the 0.07 expected by chance. The 28% variance accounted for is barely beyond the expected level of random variation of 25% (11 *K).* In multiple factor analysis, the ratio of variance of the first multiple factor is 46% compared to 24% (a 2-to-1 rather than the minimal 3-to-1 ratio). There is a slightly lower association than expected (see Table lIA) between measures one and four (Paxson and Whyte) that loads them opposite to one another on factor 2.

Rohners' scale appears from Table lIB to be more reliable than the other measures (this is also the case for the odd-numbered subsample coded by Whyte in which the same pattern of reliabilities is replicated).[13] Why might this be so? One possible answer is the differing concerns with reliability and data quality in the four studies. One indicator of the uncertainty that coders had about these variables (even pairs of coders discussing independent codings) is the high degree of missing data. The Rohner codes have 64% missing data; the Whyte codes 32% of the odd

TABLE IIB

Reliabilities for Father — Child closeness measures.
$N = 186$ cases (93 note coded in 614)

VARIABLES Ref. Num.	Coder	INDICATORS Children's Ages	Father's Role	B Factor Loadings give: Reliability B^2	Optimal Scaling Correlation R	Optimal Scaling Scaling Weights W
54	Paxson	2—5	Father's closeness	0.34	0.32	0.34
369—70	Josephson	4—7	Sex of Parental Caretaker	0.31	0.30	0.40
991	'Rohner'	3—6	Father's Caretaking Role	0.42	0.41	0.42
614	'Whyte'	5-up	Male Caretaking Role	0.05	0.09	0.13

Variance of the factor is the sum of $B^2/K = 1.12/4 = 28\%$
Additive Scale Reliability (from raw data) = 0.63
Optimal Scale Reliability (from raw data) = 0.66

numbered societies coded; the Paxson codes 19%; and the Josephson 10% for boys ratings and 6% for girls. The Rohners' coders were cautioned against making inferential ratings. But this does not explain the low reliability of Whyte's (relatively cautious) ratings and the higher reliability of Josephson's (more complete) codings. A closer examination of the techniques used in coding, however, may provide further indicators of differential reliabilities.

Rohner and Rohner (1982; Rohner 1981: 103) were extremely careful about data quality. Their coders kept independent "uncorrected" codes for each coder throughout the coding process, although coders met at regular intervals to discuss and compare codes and resolve their differences, variable by variable and society by society, so that each coder could produce a "corrected" code; since these still differed, they were summed to get the final codes. The uncorrected intercoder reliability of "overall importance of fathers for children" (0.70) was a result of convergence in coding procedures guided by the project director. Of the 186 societies, 74 were coded, 28 not coded on any variable (lacking of information in the sources), and 84 were coded on some variables but not on the overall importance of father as a caretaker (for lack of information in the sources).

Barry and Paxson (1971: 168) report three levels of missing data or doubtful coding: (1) information lacking or excessively meager or conflicting (coded'.' in 30 cases); (2) highly doubtful or uncertain ratings, when the information is meager or contradictory or when the code is based on a single instance or a weak inference [coded in brackets in 16 cases]; and (3) ratings somewhat doubtful or a strong inference when the information is substantial but ambiguous or not directly pertinent to the code (coded in parentheses in 48 cases). They found coding difficulties of one sort or another in 94 cases, or half the sample.

Barry, Josephson, Lauer, and Marshall (1977), report missing data only in terms of "ethnographic information insufficient" (7 cases), and (5 cases) "not applicable" because the principal caretaker at this age was exclusively non-parental. They do not report on the reliability of individual codes, nor do they discuss how they achieved a greater degree of consensus coding on these codes compared to Paxson's codes.

The two sets of codes by Barry, then, indicate concern with data quality, but willingness to make strong or weak inferences. In the first study (Paxson) the authors kept track of these inferences, while in the second study (Josephson) they did not, and were willing to code an even higher proportion of cases. There is little difference in Table lIB, however, in the reliabilities of these two studies, with the later study (Josephson) having a slight edge.

Whyte's (1978: 337) two coders used "no information" quite frequently (for 25 of the 93 societies coded): they "were instructed not to 'squeeze'

information into existing categories in the case of ambiguity, but to use a 'no information' or comparable category provided with each code." It is not clear whether the low correlations (leading to low reliability estimates) are due to relatively greater unfamiliarity with ethnographic literature, or to a conceptualization of the variable referring to "male" caretakers rather than fathers in particular.

Are the relatively low reliabilities observed for some of these variables due to less reliable inferences or to differences in the conceptualization of the variables? Evidence on this question may be gleaned by (1) using Paxson's "weakness of inference" codes in further investigating sources of reliability and (2) estimating reliabilities for a restricted sample of cases consisting of only those which were coded on all four projects.

Paxson's coding of the "degree of inference" involved in coding father's proximity (variable 54) is correlated with a distant father *(r* = 0.27) and with the mother as the major caretaker (variable 369, *r* = 0.16). These correlations are probably due to an expectation bias when making inferences on the part of Barry's coders toward "mothering" (with father distant) as the normal caretaking arrangement (a "Mediterranean" family coding bias?). This bias also affects the composite factor scale *(r* = 0.22).

There is a more serious problem, however, with Whyte's variable (614). Father's authority over children (614) is positively correlated *(r* = 0.28) with less proximity (variable 54; father more distant) for the low inference cases, but *negatively* correlated *(r* = -0.33) for cases rated by Paxson as involving inference. Assuming Paxson's codes to be the more accurate, this could be a result of Whyte's coders "guessing" from father's greater proximity that he has greater authority (when the opposite is more likely to be the case in fact), and "guessing" from lack of proximity that the mother has more authority (when the opposite is again more likely). This might be a "Yankee" type of coder expectation bias, as opposed to a "Mediterranean" type bias of Barry's coders.

Is there evidence that Paxson's coders are more accurate, as against Whyte's? First, we have the evidence of Table **IIB,** suggesting that Paxson's codes are more reliable. Stronger evidence, however, is that: (1) Paxson's codes correlate strongly with Rohner's *(r* = 0.51) for the no inference cases, but not for the inferential cases *(r* = 0.14); while (2) Whyte's codes are uncorrelated with Rohner's for the no inference cases *(r* = 0.03) but *negatively* correlated *(r* = -0.54) for the inferential cases. The latter evidence replicates the negative correlation with Paxson's variable, and supports the view that Whyte's coders are projecting expectations about father's role when the ethnographies are ambiguous.

Before rejecting Whyte's code because of systematic measurement bias, one more replication may be tried. The 24 cases coded by all four projects are presumably those societies with the best ethnographic information on father - child interaction. Seventy-three percent of these cases are for

ethnographic time periods in the twentieth century, compared to 44% in the remainder of the sample (gamma = 0.38, $P < 0.001$).[14] Table ill presents the results of reliability analysis for these 24 cases. The correlations and reliabilities improve (for a composite scale reliability of 0.71 compared to an original 0.66), but Paxson's and Whyte's variables show the highest reliabilities. All three tests of one-factor structure are satisfied: in addition to factor scores accurately reproducing the correlation matrix (actual standard deviation of $0.02 < 0.20$ expected), the variance on the first factor is 33%, and the variance ratio to the second factor is 2.63, nearly 3-to-1. Whyte's codes (and Paxson's), in this context, are no longer so unreliable.

In short, by restricting the analysis in Table ill to higher quality cases, or eliminating any cases with missing data, differences in reliabilities of the codes are removed insofar as they are due to difficulties in making inferences for ambiguous cases. The reliability of the Rohners' code has gone down if only because much of the random error has been eliminated in the otherwise less reliable codes.

This example makes clear the usefulness of the idea of a measurement domain defined by overlap in meaning that centers on the main measurement concept, but allows unique variation as well. Items that are similar but not identical in content differ not only in random error (errors of sampling or coder "guessing"), but in unique systematic variance or substantive differences, as well as potential measurement bias, such as have been identified. But with data that fit the one-factor model, measurements that are "unique" to each variable come out as self-cancelling measurement errors. A corollary is that if variables were examined each against independent coding replication of *exactly the same* coding definitions and procedures, their reliabilities might be considerably higher.

One might have concluded from inspection of the definitions of variables and their reliabilities in Table IIB that two variables - Josephson's and the Rohners' - had the greatest overlap in meaning (both refer to "fathers" caretaking children of approximately the same age, in contrast to "mothers." Paxson's variable relates to father's "closeness" but neither in contrast to the mother, nor specifically in terms of caretaking. The Whyte variable relates to caretaking by "males"). Although the variables do have different substantive meanings, the results in Table IIB and III do not support a relationship between greater reliability and greater overlap in meaning. It is not clear, however, whether coding procedures or substantive differences in definition of the variables (e.g., a distinct systematic dimension of measurement) are the cause of the different reliabilities in this case.

In contrast to the rather simplistic notion often used in cross-cultural research of simply choosing what appears to be the best variable, the example illustrates the use of a variety of techniques for improving

TABLE III

Reliabilities for Father — Child closeness measures.
$N = 24$ cases with no missing data codes in any of the four studies

VARIABLES Ref. Num.	Coder	INDICATORS Children's Ages	Father's Role	B Factor Loadings give: Reliability B^2	Optimal Scaling Correlation R	Optimal Scaling Scaling Weights W
54	Paxson	2—5	Father's Closeness	0.40	0.40	0.48
369—70	Josephson	4—7	Sex of Parental Caretaker	0.27	0.27	0.41
991	'Rohner'	3—6	Father's Caretaking Role	0.23	0.24	0.38
614	'Whyte'	5-up	Male Caretaking Role	0.40	0.40	0.48

Variance of the factor is the sum of $B^2/K = 1.30/4 = 33\%$
Additive Scale Reliability (from raw data) = 0.68
Optimal Scale Reliability (from raw data) = 0.71
Average Correlation of 0.32 gives a simple average composite scale (Cronbach) minimal reliability of 0.65.

measurement, starting from the concept of multiple measurement of the same domain, conceptually defined (Le., not restricted to testing intercoder reliabilities, or attempts to code exactly the same variable).

This examination has shown four measures which initially fail to provide an adequate degree of reliability, which fail the one-factor test, and which cannot be combined into a summary scale because of problems with individual items. In particular, there is evidence that one of the variables, that of Whyte, has an unacceptable degree of measurement bias. The bias appears to be that of coder expectations in cases where the ethnographies are ambiguous (e.g., see Bradley 1989): if the father is in physical proximity to the child, he is assumed to have authority and discipline over the child; while if the father is distant, the mother is assumed to have authority. This bias is rather severe, and disappears only when, by examining only the highest quality ethnographies (as evidenced by the lack of missing data codes), we restrict the sample to such a small size *(N* = 24) as to limit its usefulness. This variable, then, should probably be eliminated from consideration in evaluating reliability and in summary scale construction. We have also shown that another of the variables (Paxson's) appears to have a moderate coding bias towards assuming, when ethnographies are ambiguous, that "mothering" (father distant) is the normal caretaker arrangement.

The bias in Paxson's variable can be eliminated by restricting the sample to cases *(N* = 86; 78 coded on the other variables) where Paxson did not report using inferences in coding. Table IV shows the reliability analysis, for the three valid variables, for this restricted but sizeable sample. The reliabilities, factor structure, and summary scale reliability are acceptable for further use of this composite scale - or the individual variables - but only for the restricted sample. The composite scale reliability is 0.79, close to the level of 0.80 recommended by Nunnally.[15]

ISSUES OF VALIDITY IN THE FATHER - CHILD MEASUREMENT DOMAIN

The fact that there is considerable variation (random error) in correlations due to sampling as well as random measurement error is one reason for careful tests of the single-factor model in reliability theory. The example of father - child interaction variables provides good fit to a single-factor model for the father - child interaction codes when they are done on high-quality ethnographies, but poor fit when the ethnographies require inferences. There is considerable systematic measurement error representing different coding procedures and coder expectations in the high-inference ethnographic cases (see Bradley 1989 for a comparable case). But even if goodness-of-fit of a one-factor structure is established, *factor*

TABLE IV

Reliabilities for Father — Child closeness measures.
$N = 86$ (78) cases where Paxson did not require inferences in coding

VARIABLES		INDICATORS		B Factor Loadings give:	Optimal Scaling	
Ref. Num.	Coder	Children's Ages	Father's Role	Reliability B^2	Correlation R	Scaling Weights W
54	Paxson	2—5	Father's Closeness	0.56	0.39	0.59
369—70	Josephson	4—7	Sex of Parental Caretaker	0.38	0.49	0.35
991	'Rohner'	3—6	Father's Caretaking Role	0.47	0.55	0.45

Variance of the factor is the sum of $B^2/K = 1.41/3 = 47\% > 33\%$
Additive Scale Reliability (from raw data) = 0.76
Optimal Scale Reliability (from raw data) = 0.79
Average Correlation of 0.46 gives a simple average composite scale (Cronbach) minimal reliability of 0.72.

scores do not provide their own validity (Schuessler 1971: 118). *They must be validated by external evidence.*

To examine construct validity issues in terms of how well variables predict an established result, a useful example is the correlation between father - child interaction variables and beliefs in high gods, reported by Jester (1989). Jester's argument follows Terry (1971), who found an association between nurturant socialization and the absence of monotheism. Jester (1989: 6) reformulates Freud's argument for projection of infantile childhood relationships with the father onto beliefs about gods:

> The nature of belief in a high god is that of a child searching for a father, not that of a child with a father. As Terry has pointed out, a nurturant socialization is inversely related to monotheism. And, in line with (Freud's] metaphor of personalizing natural forces in order to be able to placate them, the father, projected as a god, should also be remote from the emotional life of the child. Freudian theory should be restated to point out that it is, in fact, the lack of a human father, more precisely, the lack of a nurturant, supportive relationship between the father and the child, that establishes the context for belief in gods.

According to this hypothesis, belief in nurturant and supportive gods express an unconscious longing for a nurturant and supportive father, in the absence of such a relationship with the actual father.

Correlations between high gods and father - child interactions (standardized to measure "closeness") for the total sample are in the direction predicted by Jester and Terry but rather small for the Paxson, Josephson, and Rohner variables *(R* = -0.20, -0.20, -0.16), and opposite for the Whyte variable *(r* = 0.26). The correlation for the four-variable composite score is -0.18. With the *(N* = 86) sample of no-inference cases, these correlations improve dramatically for the Paxson and Josephson codes *(r* = -0.34, -0.31) and three-variable composite score *(r* = -0.31). For the Whyte variable, correlations opposite to that predicted, in both samples, are somewhat more consistent with the finding of coding bias than that of substantive differences in code definition.

If we did not have coded data from Paxson on the extent to which coding the father - child relationship required inference, could we still find an appropriate test of the validity of correlations with more reliable segments of the sample? Can we use intercoder disagreements to identify a subset of the most reliable cases and improve the correlation between individual father - child variables the high god variables? First, we examine the correlation between each of the four father - child variables and high gods using the extent of missing data as a control, and attempt to test the validity of apparent coding biases. Second, we examine the correlations with high gods using the Paxson rating of degree of inference as a control. Third, we summarize the data and examine the correlations with high gods using the variable of intercode disagreement (reliability-by-case analysis) as a control.

The correlation (ignoring sign) between Paxson's code and high gods is 0.20 in the full sample *(N* = 137 discounting missing data). Table V rearranges the ordering of the high gods scale to test Jester's (1989) modification of the Freudian hypothesis that predicts a "distant" father associated with neutral or supportive gods. The correlation between Paxson's code and the high gods scale increases to 0.37 in the subsample of cases coded in at least three of the four studies *(N* = 85), but decreases to 0.03 in the subsample where only one or two of the studies are coded. These results support both the hypothesis and the validity of the Paxson variable.

TABLE V

Distant father correlated with High Gods by extent of missing data.

A - Three or Four codes coded	HIGH GODS			
N=85				
FATHER'S ROLE	Absent	Nonsupportive	Neutral	Supportive
1 No Proximity				2
2 Rarely Close	1		5	3
3 Occas. Close	10	1	9	7
4 Freq. Close	19	6	7	7
5 Companionship	6		1	1

r= -0.37, 14%ofvariance,p < 0.001

One or Two of four codes coded

B - Weak or strong inferences used in Paxson Codes				
N=52				
FATHER'S ROLE	Absent	Nonsupportive	Neutral	Supportive
1 No Proximity		1		
2 Rarely Close	2	1	3	
3 Occas. Close	6	1	5	5
4 Freq. Close	12	1	7	7
5 Companionship	1			

r = -0.03, 0% of variance, *p* = n.s.

Similarly, the correlation between Josephson's code and the high gods scale increases to 0.39 in the subsample of cases coded in at least three of the four studies *(N* = 85), but decreases to 0.03 in the subsample where only one or two of the studies are coded. Rohner's codes are modestly correlated with high gods *(r* = 0.19) in the first subsample, and too small to correlate *(N* = 3) for the second.

Whyte's codes, however, while they show a drop, from 0.28 to 0.09, are correlated in the opposite direction to that predicted by the theory, even for the better quality subsample! This is consistent with the evidence for systematic bias in Whyte's codes, namely, of projecting expectations about father's role when the ethnographies are ambiguous. In the "best" sub-

sample, defined by no missing data for any variable, Whyte's variable is uncorrelated with high gods *(r* = 0.05).

Is there external evidence to validate the claim that Whyte's codes are systematically biased? Examining correlates with data quality control variables (Rohner, Berg and Rohner 1982), we find that fewer months of fieldwork, fewer informants, and greater use of male informants correlate (0.32, 0.29 and 0.28, respectively) with higher estimates of male authority over the discipline and caretaking of children. This is consistent with a possible pattern of "Yankee" bias discussed above, in Whyte's codes.

Rohner's coders, on the other hand, tend to see fathers as more important in caretaking children where ethnographers have less language familiarity *(r* = 0.28; possibly a failure to discount bias of missionary reports), but mothers as more important with fewer pages on child rearing, societies rated lower in overall data quality, or male ethnographers *(r* = 0.24, 0.19, 0.20; perhaps a Mediterranean "mothering" bias shared with Barry's coders). The Paxson coders show evidence of this latter type of bias where there are relatively few total pages of ethnography *(r* = 0.23), and the Josephson coders (who seem to read more selectively) where there are relatively few pages on child rearing *(r* = 0.21).

Table VI shows an increase in the correlation between Paxson's code

TABLE VI

Distant father correlated with High Gods for Paxson's quality data.

A - No Inferences in Paxson Codes	HIGH GODS			
N = 86 (78 with data)				
FATHER'S ROLE	Absent	Nonsupportive	Neutral	Supportive
1 No Proximity				1
2 Rarely Close		1	3	2
3 Occas. Close	5	1	4	6
4 Freq. Close	23	5	7	12
5 Companionship	7		1	

r=-0.38, 14% of variance, p < 0.001

B - Weak or strong inferences used in Paxson Codes				
N = 69 (59 with data)				
FATHER'S ROLE	Absent	Nonsupportive	Neutral	Supportive
1 No Proximity		1		1
2 Rarely Close	3		5	1
3 Occas. Close	11	1	10	6
4 Freq. Close	8	2	7	2
5 Companionship	8	2		1

r = -0.07, 1 % of variance, *p* = n.s.

and the high gods scale, to 0.38 in the subsample of cases where Paxson reported that no inferences were made in coding, *(N* = 86, 78 no missing data), but decreases to 0.07 in the subsample where strong inferences were used *(N* = 64, 59 no missing data). These results again support the hypothesis and even more strongly the validity of the Paxson variable and the coding of the inference scale. Table VII shows a substantial and parallel increase to 0.42 from 0.22 for Josephson's codes.

Table VITI shows evidence of bias in Whyte's codes, with a high correlation *(r* = 0.40) opposite to the direct predicted in the *more* inferential cases, and a low correlation *(r* = 0.19) for the no inference cases.

The analysis above benefits from the fact that we have external evidence about the degree of inference used in coding, as rated by Paxson, and we can take the number of missing data codes as a proxy measure for ambiguous or poorly described cases. What if we had neither of these sources of information about bias? Could we tell data quality from the internal evidence of the codes themselves, i.e., from the extent of their internal disagreements? This could be an extremely important result for a reliability theory that applies to the cases coded as well as the variables.

A summary of results of testing the high gods correlation is given **in**

TABLE VII

Fa's caretaking correlated with High Gods for Paxson's quality data (Josephson's variable).

A - No Inferences in Paxson Codes	HIGH GODS			
N = 86 (69 with data)				
PARENTAL CARETAKERS	Absent	Nonsupportive	Neutral	Supportive
1 Father Exclusively	1			
2 Father Predominant	2	1		
3 Equal	13	1	2	2
4 Mother Predominant	61	4	3	3
5 Mother Exclusively	11	1	13	13

r = 0.42, 18% of variance, $p < 0.001$

B - Weak or strong inferences used in Paxson Codes				
N = 69 (50 with data)				
FATHER'S ROLE	Absent	Nonsupportive	Neutral	Supportive
1 Father Exclusively				1
2 Father Predominant	1			
3 Equal	3	1		1
4 Mother Predominant	4		7	
5 Mother Exclusively	13	3	13	9

r = 0.22,4% of variance, *p* = n.s.

TABLE VIII
Fa's authority correlated with High Gods for Paxson's quality data.

A - Whyte's variable	HIGH GODS			
No Inferences in Paxson Codes				
N = 86 (31 with data)				
DISCIPLINE OF CHILDREN	Absent	Nonsupportive	Neutral	Supportive
4 More Mother	1			
3 About Equal	9	1	3	3
2 More Father	3			4
1 Father	1		1	
r =(-)0.19, 4% of variance, p < n.s.				
B - Weak or strong inferences used in Paxson Codes				
N = 64 (29 with data)				
FATHER'S ROLE	Absent	Nonsupportive	Neutral	Supportive
4 More Mother	2	1	4	
3 About Equal	5	2	5	1
2 More Father	1		1	2
1 Father			3	2
r =(-)0.40,16% of variance, p < 0.001				

Table IX. The first three rows give correlations with the original high gods code (absent, not active, non supportive, supportive). The next seven rows give correlations with the high gods scale with the two middle categories reversed (absent, nonsupportive, neutral, supportive) as a better test of the hypothesis. Correlations with Paxson's and Josephson's variables (0.37 to 0.45), for the high-quality samples (no inference or low missing data), are nearly one and a half times to double the magnitude of correlations for the total sample (0.20 to 0.30). The magnitude of the high-quality sample correlations is especially high considering the estimated reliabilities of these two variables (0.56, 0.38), and the correction for attenuation of correlations (see note 3, Equation (2'».

It is significant in Table IX that the best data quality control variable discriminating high from low correlations with the dependent variable of high gods is the internal scale agreement (case reliability) rating. This lends substantial credibility to reliability-by-case analysis, identifying unreliable ratings from internal evidence of disagreements among multiple observations.

Does forming a composite scale lead to a better prediction about high gods? This does not automatically follow from factor scale analysis, since items combined in a scale may still have unique systematic variances that contribute individually to predicting an independent variable, but do not contribute to greater scale reliability. Multiple regression for the no

TABLE IX
Tests of Association between Father — Child Variables and High Gods.

CORRELATIONS WITH HIGH GODS	FATHER-CHILD VARIABLES					
	Paxson	Josephson	Rohner	Whyte	Composite 4-Vars	Scale 3-Vars
High Gods Original Code						
Total Sample	−0.20	−0.20	−0.16	[0.26]	−0.18	N.A.
No Inference Sample ($N = 86$)	−0.34	−0.31	−0.17	[0.16]	N.A.	−0.31
High Inference Sample (69)	−0.50.	−0.20.	−0.18.	[0.36]	N.A.	−0.12.
High Gods Recoded (2/3)						
Total Sample	−0.25	−0.30	−0.22	[0.27]	N.A.	−0.30
Scale Agreement Sample (108)	−0.45	−0.39	−0.31	[0.12]	N.A.	−0.39
Scale Disagreement ($N = 75$)	−0.09	−0.14	−0.11	[0.50]	N.A.	−0.12
No Inference Sample ($N = 86$)	−0.38	−0.42	−0.24	[0.19]	N.A.	−0.38
High Inference Sample (69)	−0.07	−0.22	−0.20	[0.40]	N.A.	−0.11.
Low Missing Data ($N = 85$)	−0.37	−0.39	−0.19	[0.28]	N.A.	−0.36.
High Missing Data ($N = 100$)	−0.03	−0.03	$N = 3$	[0.08]	N.A.	−0.03.

inference sample, for example, shows that both Paxson's and Josephson's variables contribute to predicting high gods at a multiple *r* that is higher *(r* = 0.44) than the composite scale correlation (0.38).

Thus, while we have used the one-factor model to estimate the reliabilities of these variables, it need not follow that the factor-score model will be a more valid measure of an independent variable for a specific hypothesis than the combination of variables using linear regression.

PRACTICAL IMPLICATIONS OF RELIABILITY THEORY

1. *Assumptions*

Reliability theory can be used to estimate reliabilities of composite scales for any multiple measurement sampling of a domain (including sets of items that are not strictly independent). The domain sampling model allows much more than simple inter-coder reliabilities. The combined reliability of heterogeneous variables with common conceptual overlap, for example, is easily assessed.

Individual item reliabilities can be estimated only if the assumptions of the single-factor model hold, as may be tested by single-factor methods. The assumptions that must be matched in practical application of individual item reliability theory are that the common measurement component is a valid measure of the concept (i.e., that one has selected variables that match the concept), and extraneous measurement errors of each variable are uncorrelated (random elements are uncorrelated by definition; systematic errors mayor may not have common variance). As the example has illustrated, this issue is trickier than it seems, and may involve extensive validity testing.

Failures of the one-factor model or in tests for measurement bias may invalidate the use of the model for particular variables or sets of variables. If single-factor assumptions do not hold, individual reliabilities cannot be estimated, but composite scale reliabilities can still be assessed.

The single-factor model discussed here can also be used for scaling related measures that are not strictly independent, but here the individual reliability interpretation does not apply. **In** this case, the use of principal component analysis (PCA) is a more general method of unidimensional (as well as multidimensional) scaling (Weller and Romney, forthcoming). Factor loadings in PCA should not be interpreted as reliabilities.

Although it is necessary in practice to apply individual-item reliability theory only to "strictly" independent measures of the same concept, there are a number of ways to achieve such independence. What varies from measure to measure may depend on context - variation in coders on the same project, related measures by the same coders, or different but similar

measures from different projects. Individual item reliability is relative not absolute.

2. *Sampling Domains*

Items from a sampling domain are not, in general, identically conceptualized. Measures by different coders or observers, or from different materials or questions, often differ markedly in their reliability in respect to a common measurement component, and the different degrees of "unreliability" in each measure may reflect not simply random error but different and uncorrelated sources of systematic error (each a measure of something else besides a shared component), even assuming statistical support for a single-factor model.

Results of the program and methods described here for one-factor scaling can also be used for reliability interpretations across several types of domains. For example, with different sets of measures of a domain, each of which is done by independent coders or research teams (coding the same societies, for example), the domain of variables selected is crosscut by a domain of methodological variability across different projects or coders. Dissimilar measures by the same coder, for example, mayor may not correlate more highly than similar measures by different projects or coders. Another domain might separate codes according to differences in the sources used in the coding. Reliability coefficients obtained from the single-factor analysis could be subjected to analysis of variance by type of domain. For such problems, the study of reliability under different conditions grades into the problem of validity (see Campbell and Fiske 1959).

From various perspectives, then, reliability is not an absolute, a property of a variable or scale in and of itself, but of how the scale or variable was derived in comparison with other scales or variables of similar derivation. This is consistent with Tryon's (1957) approach to the "comparable construct" measurement of reliability.

3. *Few vs. Many Measures*

The general point that more are better than fewer measurements is not valid for just any combined measures, but is valid (even when some items have low reliabilities) with the use of optimally weighted factor scales, *provided* that all measures contain only random error and uncorrelated extraneous measurement components. Optimal reliability scores give low (or zero) weights to the more unreliable variables.

By averaging items of nearly equal individual reliabilities, composite measurement scales can achieve any desired degree of reliability. If there is high variance in reliabilities, where one or a few variables have reliabilities far above the rest, there may be little loss in neglecting the less

reliable variables (especially those with reliabilities below 0.1 or single-factor loadings below 0.3).

Only by the study of individual reliabilities can one decide objectively whether taking only the most reliable measurement(s) is a reasonable practical strategy, as opposed to the "multiple measurement" approach of combined measurement.

4. *Direct vs. Indirect Measures*

Indirect measurement involves unobserved constructs inferred from other indicators or symptoms, such as "mental abilities" inferred from verbal, written or behavioral responses. The issues with such measures are: (1) taken individually, they may have relatively low reliabilities (but as we have seen, composite scales that incorporate many measures of uniformly low reliability may be designed for any desired level of reliability), and (2)
. they may contain extraneous measurement components that affect validity (but selection of variables for non-overlapping extraneous components of measurement, and tests of the one-factor model act as a guard against extraneous measurement, and can secure validity).

Although one would think that direct measures having greater face validity would be more reliable than indirect measurement, they are no guarantee of more reliable measurement. The choice of single direct measures thought to "best" represent a domain of measurement is no substitute for the analysis, where possible, of individual reliability by multiple indicators.

ESSENTIALISM VS. VARIATIONISM

In addition to straightforward assessment of composite scale reliability, reliability theory uses measurement variation as a means of testing the assumptions of the one-factor model, testing for systematic bias, estimating reliabilities of individual items, and creating optimal composite scales. What then, is the reason why the assessment of reliabilities is so rarely applied in anthropology, including comparative or (beyond inter-rater correlations) cross-cultural studies? One reason may be a "distrust" of composite scales, possibly related to an "essentialist" preference for single variables taken to be the "best" representative of a given measurement concept. Three points of contrast characterize the variationist approach to reliability emphasized here as opposed to an "essentialist" approach which assumes that there is a single "best" measure.

The first contrast concerns the value of composite scales. The essentialist approach tends to reject composite scales in favor of single

measures that are often argued to be "best" on the basis of face validity. With the variationist approach of factor weighting, as items are added to a measurement scale, although they may be less reliable, composite scale reliability increases so long as there is no systematic measurement error that is not unique to this variable. Composite scales, however, are not self--validating. Thus, either assumption requires further validation.

The second contrast concerns differences, in the interpretability of measures from the two viewpoints. The essentialist will argue that the best single measure is more interpretable. The variationist will argue that even the best single measure is more contaminated with unknown sources of systematic error than the composite measure, and the best single measure is therefore less interpretable in terms of measurement concepts.

The third point has to do with the evaluation of the variables. The essentialist may tend to judge certain variables as "intrinsically" bad because of their low correlations or estimated reliabilities. The variationist may note that measures of reliability are affected by the measurement context. Even variables that are relatively low in reliability, provided they fit the one-factor model, may have high reliability in another context.[16] Reliability is a relative concept, and its measurement varies with the context. Variables are neither "intrinsically" reliable or unreliable, and measures of reliability can best be assessed as a range of values (or mean and confidence limits). While reliability estimates are not independent of the measurement context, neither are they arbitrary.

It is preferable, of course, to have variables of high- rather than low--reliability. This is not always possible. Measures of father-child closeness are obviously a domain where this has not been achieved for the total sample for ethnographic cases. Yet the example has shown that it is still possible to have composite scale reliabilities as high as 0.79 for restricted high-quality subsamples of substantial size.

PROBLEMS OF VALIDITY: SUMMARY

In sampling one or many measures from a domain, "The major source of measurement error is because of the sampling of content" (Nunnally 1967: p. 211). This applies not only to random error, but to the introduction of systematic error, and threats to the validity of measurement.

Reliability is a necessary but not a sufficient condition of validity. The systematic component of an "operational" measure - such as the first component in factor analysis - may not be perfectly aligned with the measurement concept. The alignment procedures are to (1) test for and/or eliminate extraneous components of measurement that are neither random nor correlated with the first factor (the systematic measurement component in the one-factor' model), and (2) test or justify the assumption that

the first systematic component in the measurement is aligned with the measurement concept by the choice or sampling of variables. These procedures, however, do not guarantee a fit between the systematic measurement component and valid measurement of the concept. That is, empirical clustering of measures is a good strategy for selecting certain common elements, but the question of how they relate to the original concept (or what to call this common variance) remains a conceptual or theoretical problem.

Sampling variables from a domain of measurement is often done on the basis of face validity, and again offers no guarantee of valid measurement. In the area of intelligence testing, for example, it is now more clear than ever that what so many people labeled intelligence was in great part acculturation (and even literacy). The problem with intelligence testing was that proper tests of measurement bias were not applied early on.

Many types of systematic bias in measurement can be detected by creating data quality variables that will correlate with measured variables if they are sources of bias (Naroll 1962), assuming that different degrees of bias are evidenced in different cases. The most serious problem, of course, is if all measures are biased uniformly and in the same direction because of a problem in conceptualizing the theory or operationalizing the variables of the theory. In such cases, improvements in theory and conceptualization, leading to new types of measures, are necessary to identifying problems of validity.

Construct validity is an approach to convergent measurement guided by theory: take the most divergent predictions generated by a theory that specifies measurement operations in different contexts, and test whether the predictions are replicated within the tolerance of random error.[17] Replications under divergent conditions are central to establishing the validity of measurement, in addition to testing specific hypotheses.

CONCLUSION

Procedures for the study of reliabilities apply to any domain with multiple measurements of the same concept. Composite scale reliability can be assessed from two or more measures whether or not they are independent. To assess reliabilities of individual items, the measures must number three or more and be strictly independent. From reliability analysis of individual variables, optimal reliability scores can be constructed, and an ORS index (constructed by the method of single-factor weights) will have a reliability at least as great as any of its component measures. Application of the single-factor model to the question of whether the variables do indeed have a single measurement component provides crucial information on how well the measurements fit a common conceptualization of the

variables, thereby providing an evaluation of one aspect of measurement validity. Estimates of reliability by individual cases (rather than variables) can be done to establish subsamples with which to replicate correlational tests of hypotheses on the assumption that valid correlations will be higher in more reliably measured samples.

Where reliability analysis of individual variables can be validated within the single-factor model, it is argued that optimal factor scores are both more reliable and more interpretable than the best single-variable measures. Evaluation of single-factors measurement models is also a critical part of measurement validation and detection of potential measurement biases.

While measurement validity is perhaps the single most important issue in empirical research, factor scales are not self-validating even if they fit a single-factor model. Factor scales, as well as individual variables, must also be externally validated, and tested against external criteria for potential sources of systematic measurement error, or bias. Reliability analysis can help in identifying some of the potential sources of systematic variation that differ from the principal measurement component. Further investigation of thy sources of these extraneous measurement components may be needed to identify such measurement biases, however. Analysis of reliability is thus a necessary but not sufficient component of measurement validation.

The present application of the theory of reliability estimation to the problem of measuring father - child interaction illustrated some of the types of solutions to problems that arise when variables are based on high degrees of inference, with potentially serious problems of measurement bias. Four strictly independent measures for father-child closeness were analyzed. The first analysis showed that these variables did not fit the one-factor model, and thus the one-factor model could not be validly used to estimate their reliabilities. One might have supposed at this point that the four variables in fact measured very different 'concepts of father-child interaction. Alternatively, problems of inferences made by coders in cases where the ethnographies were ambiguous could have led to substantially different systematic errors in the variables. Restricting the sample to cases where none of the studies reported missing data showed that the one-factor model did apply to this higher-quality subsample. A variable coding the degree of inference used in one of the codes, used as a control variable, showed strong evidence that one of the four main variables suffered from measurement bias, probably from "guessing" from father's greater proximity that he had greater authority or discipline over the children than the mother, when the opposite is in fact normally the case. The fact that all four codes fit the one-factor model in the subsample with no missing data (presumably the highest ethnographic quality), supported the notion that the problem with the fourth code was one of measurement

bias, rather than one of measuring a different dimension of father-child interaction. Strong measurement bias would justify eliminating the fourth variable. Finally, the three remaining variables, for the low-inference or high data quality subsample, fit the one-factor model, show much better individual reliabilities, and combined to produce a composite score with sufficiently high reliabilities (0.79) for use in relatively accurate testing of correlational hypotheses (correlations are significantly attenuated by reliabilities much lower than 0.8).

Further tests of validation confirm the hypothesis of bias in one of the variables, and confirm the expectation that, with a significant underlying correlation, tests with a high-quality subsample will give higher estimates than tests with the low-quality subsample. Significantly, the reliability-by-case variable showed the best discrimination of reliability in subsamples, as determined by testing correlation with a criterion variable.

The hypothesis used in exemplifying the external validity tests was Jester's (1989) modification of Freud's expectation of an association between belief in high gods and the remoteness of father and child. Besides helping to validate three of the four father-child variables, and to indicate measurement bias in the fourth, this hypothesis was supported.

Thus, beginning with high-inference variables, each of which is faulty in different ways (one more seriously than the rest), strategies for assessing reliability and measurement bias led to identification of a set of variables, a high-quality subsample, and a multiple measure composite scale defined on this subsample, all of which pass the one-factor test as a requisite to estimating item reliabilities, and which pass various other tests of validity. The final scale has a reliability (0.79) comparable to many low-inference variables.[18] Valid hypothesis tests, then, can be achieved in this domain.

While exemplified only for cross-cultural research, identical procedures can be used with equal effect for the study of reliabilities in case-study data from particular field studies. The procedures and computer program for reliability analysis that have been developed and used here should prove equally useful in ethnographic and ethnological research. It should be stressed, however, that strict independence among the measures
given the measurement context - is a prerequisite to individual item reliability analysis, and that the use of composite factor scores (or simple additive scores) is not recommended without external tests of measurement validity.

The theory of reliability applies more generally to the study of culture, and includes assessment of informant accuracy (Romney, Weller and Batchelder 1986). When combined with assessment of the reliability of variables, as in the present context, reliability theory provides a powerful framework for assessing and enhancing the validity of anthropological research.
The development of an adequate practice and theoretical understanding

of reliability analysis will take anthropology a step closer to both a more humane and scientific worldview wherein the foundations of the discipline can be reintegrated. Anthropologists have historically been rightfully distrustful about the claims of other anthropologists to have correctly identified "essential" phenomenon or descriptions, yet they have lacked a set of practices that would allow them to evaluate one another's data. Those who hold essentialist views, however, are likely to distrust all data they did not generate themselves. The conventional distrust of cross-cultural analysis among anthropologists, for example, may exemplify a widespread essentialist bias among anthropologists. It is all the more important, then, to establish clear-cut procedures and research practices for the evaluation of reliability, and to supplant "essentialist" attitudes with variationist practices. The theory of reliability is thus potentially unifying for the field. By focusing on different and multiple ways of measuring the same thing, and upgrading reliability through multiple measurement, reliability analysis provides an approach to the problem of data quality that is quite different than essentialist dismissals either of comparative research in general or of other people's fieldwork in particular. Ethnographic and coded data are neither intrinsically reliable or unreliable. Reliability depends on identifying contexts in which multiple independent measures can be made of the same concept Creative thinking about measurement can improve reliability, both in ethnography and comparative research. Cooperativity, in looking at the same things, independently, is productive in the realm of reliability of measurement, as elsewhere. In these ways, we help to create a part of a shared discourse, potentially both scientific and humanistic.

ACKNOWLEDGEMENT

I thank A. Kimball Romney for his help in explicating the theory of reliability, he and Roy D'Andrade for readings of the initial manuscript, Cindy Mitchell for providing research materials, and the following readers of a revised manuscript for very helpful comments: Pertti Pelto, Marc Ross, Jeffrey Rouder, Carmella Moore, Thomas Schweizer, and Ronald Cohen. I am particularly indebted to an anonymous reviewer for reference to the work of Robert Tryon, and for suggestions as to reformulation of the reliability presentation, including part of the explication of definition 3. This research was supported by NSF grants BNS-8507685 (World Systems and Ethnological Theory, co-PI Michael Burton), BNS-8718769 (Assessment of Long Term Development in Zambia, co-PIs Thayer Scudder and Elizabeth Colson), and BNS-8911173 (Summer Institutes for Comparative Anthropological Research, with co-PI Carol Ember).

NOTES

1 Nunnally (1967: 175-178) proves that as the number of items approaches infinity, the correlation $r1(1 \ldots k)$ between one item and the sum or average of all items (in which random errors tend to cancel each other out) equals r_{1t}, the correlation of one item with "true score" on the common component of measurement.

2 This is also a problem in classical test theory. Lord (1980) develops an item response theory (for dichotomous items) that provides optimal weightings such that the contribution of the item to the measurement effectiveness of the total test does not depend on what other items are included in the test. He also uses latent class analysis as a more general approach than single-factor theory. Additional mathematical assumptions are required in this approach. Batchelder and Romney (1988, 1989) also use this general approach in developing cultural consensus theory. The present approach builds on Tryon's (1957) comparable constructs reliability theory, and Spearman-Brown single factor theory, without the "parallel forms" assumptions of classical test theory (e.g., Nunnally 1967).

3 Spearman's one-factor model and correction for attenuation of the true correlation between measures, with purely random measurement error (see also Nunnally 1967: 203-204, 218), is a more general expansion of Equation (2). Given two sets of measures i and j, each with known reliability, r_{ii} and r_{jj}, the true correlation r_{ij}^* between their systematic measurement components (Spearman 1904a: 90; 1904b: 253) is a function of their observed correlation (r_{ij}), divided by each measure's correlation with true score on its measurement component:

$$(2') \qquad r_{ij}^* = r_{ij}/r_{iT}r_{jT} = r_{ij}/r_{ii}^{1/2}r_{jj}^{1/2}$$

This result is true independently of what each variable operationally measures, that is, it is not dependent on valid measurement. Consequently, this equation provides a test of the one-factor model, since the estimated values may take values greater than 1 only where the single-factor model is violated.

Separate reliabilities, however, cannot be estimated for only two variables, and the means of estimating reliabilities for three variables (Equation (3» make $r_{ij}^* = 1$ true by definition in the case of three variables. Thus, the single-factor test in (2') is useful only with four or more variables. For example, say four items all correlated at 0.5. Equation (3) estimates each of their reliabilities at 0.5 also. Equation (2') will estimate the true correlation between each pair of variables at 1.0, which fits the one-factor model. The same would be 'true for any set of variables with uniform intercorrelations. However, the one-factor model may also fit cases where correlations vary as a function of different reliabilities of the items.

4 MINRES (minimum residuals) techniques solve the problem of maximally reproducing the off-diagonal elements of the correlation matrix by least-squares, without using diagonal communalities in the estimation, but obtaining communalities that are consistent with least-squares minimization.

Principal factor analysis using Hotelling's methods (Schuessler 1971: 109-114) is done iteratively, similar to the Bayesian estimation of informant reliabilities in cultural consensus theory (Romney, Weller and Batchelder 1986). This technique is convergent with the MINRES approach if communalities are reestimated at each iteration. The resultant factor loadings are estimates of true-score correlations whose products by Equation (2) are the best least-squares estimators of the original correlation matrix. The squares of these factor loadings are thus good estimates of the reliabilities of variables which may have different amounts of random error. They are also stable in that they will not change appreciably if a low-reliability variable is sampled from an otherwise high reliability domain.

Note that the type of factor loadings discussed here are Hotelling's principal factors,

not PCA or principal components analysis. PCA factor loadings do not converge with reliability estimates. On this point, see also Lord and Novick (1968: 216). Multiple-factor PCA, however, may be used as a separate test of the one-factor model.

The upper bound of item reliability, in the single-factor approach, is the multiple r^2 predicting the variable from other variable sampled in the domain. Less than optimal means of determining initial communalities, such as the use of the multiple *R2* correlation for each variable with other variables, the highest correlation in each row (which is too high for high estimates, and too low for low estimates), or "refactorization," are discussed in Schuessler (1971: 89-91).

For a general consideration of optimal weights for composite measures, see Lord and Novick (1968: 119-123). An approach to estimating reliability from ratio of true-score variance to actual variance (Definition 3) when reliabilities are not uniform is given in Lord and Novick (1968: 216-219). Lord (1980: 39-40) discusses the use of one-factor triad reliabilities in their raw form, and notes that this method is not recommended when variables are question responses that contain guessing. Correction for guessing is explicitly included in Batchelder and Romney's (1988, 1989) cultural consensus model (see also Romney, Batchelder and Weller 1987).

5 This is derived from Equations (1) and (2) by:

$$(3a) \qquad r_{ij} = (r_{ii}\, r_{jj})^{1/2}$$

so that for three measures:

$$r_{11}\, r_{22} = r_{12}^2$$
$$r_{11}\, r_{33} = r_{13}^2$$
$$r_{22}\, r_{33} = r_{23}^2$$

Solving for r_{11} :

$$(3b) \qquad r_{11} = r_{12}\, r_{13}/r_{23}$$

For more than three measures *(k* > 3), reliabilities are estimated by averaging values from multiple equations (Spearman 1904a: 90). For *N* measures:

$$(3') \qquad \hat{r}_{ii} = \text{Average } (r_{ij}\, r_{ik} / r_{jk}), \text{ for } j, k \text{ in } N, \text{ not equal to } i.$$

This estimate will be unbiased but inefficient for equations with low correlations in the denominator (e.g., < 0.1, which should be excluded in computation). This estimating technique is also unstable because of unweighted averaging: if reliabilities of several items are high, but a variable is added with very low correlations, estimates of the previous reliabilities may be substantially altered.

6 With markedly unequal reliability of individual items, the reliability of a scale that averages multiple measures mayor may not be greater than each of the individual items. Consider the case where two independent measures are perfectly correlated (1.0 reliabilities) and a third has lower correlations with each: the reliability of the average of these items is lower than that of either of the most reliable items. Weighting items by their reliabilities is an improvement over equal weighting, but does not solve the problem of reducing scale reliability if there are already items with reliabilities close to 1.0. See Lord and Novick (1968: 112-119) on Spearman-Brown reliability in the mental testing context.

7 A multiple regression model with perfect prediction of the true score of an imperfectly measured variable will have a multiple correlation *(r)* equal to the correlation between the measure and its true score. The square of the latter correlation is the reliability of the measure. Hence the multiple correlation r^2 is always less than or equal to the reliability of the dependent variable.

8 Systematic measurement error can be divided in turn into a confounded part that is correlated with the valid component of measurement and an extraneous part that is

uncorrelated with the valid component of measurement. The confounded part of measurement error is not a problem since it represents valid "proxy" measurement of the concept. The most serious part of systematic measurement error is that which is uncorrelated with the valid component of measurement.

True score estimates for each measure are computed by multiplying obtained scores by the reliability coefficient, but such estimates are biased since error terms will necessarily be correlated with the obtained scores (Nunnally 1967: 199). The magnitude of bias is inverse to reliability, in that the correlation between errors and obtained scores is the square root of (1 - reliability). Such bias is not serious where items reliabilities exceed 0.8. The problem is serious, however, when there is considerable variation in reliabilities, and some reliabilities substantially lower than 0.8.

9 For those unfamiliar with the differences between single- and multiple-factor analysis, it might be noted that in the single-factor model in Equation (2), the matrix of original correlations is reproduced from products of true-score correlations of each variable with a single measurement component. In contrast, multiple factor analysis (PCA or principal components analysis) reproduces a matrix of correlations as a sum of products of factor loadings of each variable on uncorrelated measurement components. With a single systematic measurement component, only the loadings on the first factor will reflect common measurement variance; other factor loadings will reflect variance that is unique to each variable, including random measurement error as well as unique systematic error. For multiple measurements, common systematic measurement error can thus be identified by substantial variance beyond the first factor.

In PCA multiple-factor analysis, it should also be noted that the proportion of variance accounted for by each factor times the number of variables equals the eigenvalue of the factor. The eigenvalue is also one of the solutions to the characteristic equations by which factor analysis is computed, $R\ V = e\ V$, where R is the matrix of equations, V a vector to be solved for, and e the characteristic eigenvalue constant associated with each factor, where e/N is the variance accounted for.

10 The program CORR-REL is scheduled for publication in the World Cultures electronic journal, and is also available from the author.

11 There are also variables for father - child closeness in infancy in Barry and Paxson (1971), Rohner and Rohner (1982), and Whyte (1978), but this age range is not separately coded by Barry, Josephson, Lauer, and Marshall (1977). Both sets of variables (infancy and early childhood) could not be included and preserve strict independence among the coders, which is intended to be the principal source of contrast or independence in this reliability assessment of individual variables (see assumption 3, page 111).

12 Throughout this analysis I will treat the rating scales for these variables as equal-interval measures (which most are designed to be). If these were to be treated as strictly ordinal variables, however, Tukey's (1977: 103) methods could be used to transform the proportion of cases in the distribution of the variables into a normalized interval scale. Such transformations, however, make little difference to this analysis, where the interval-level assumption is robust. On reliability estimation in the context of ordinal measurement, see Lord and Novick (1968: 214-215).

13 The correlations and reliabilities in the 93 odd-numbered subsample are slightly attenuated. The factor loadings improve the estimates of reliabilities in the sense that they converge towards the population (full sample) reliabilities. These results also fail the second and third single-factor tests.

14 These 24 cases are: Mbundu, Kikuyu, lbo, Tallensi, Pastoral Fulani, Konso, Romans, Uttar Pradesh, Garo, Andamanese, Tanala, Javanese, Toradja, Kimam, Kwoma, Siuai, Marquesans, Trukese, Japanese, Kaska, Twana, Creek, Papago, and Quiche. Note that listwise deletion is not done by the CORR-REL program used in this analysis (footnote 10), but is done by the MAPT AB data manager.

15 The first test of the one-factor model is inapplicable with three variables. The second

test is passed, with 47% of the variance on the first factor exceeding the expected 33%. The third test is passed, with the ratio of variance on the first factor (64%) exceeding that of the second (19%) by 3.3 to 1. First (PCA) factor loadings are 0.83, -0.77, and 0.80. These results are for Paxson's no-inference sample *(N* = 86). For the whole sample, however, the three variables fail the one-factor test.

16 For example, Rohner's variable (above) might be more reliable if it were assessed in the context of other variables such as measures of father-child authority (unfortunately, we have only one such independent measure, by Josephson, insufficient for reliability analysis). Note too that the Josephson measure cannot be added to the set of variables used to measure and assess reliability for father – child closeness, since (1) it is not strictly independent (in terms of the coder) of the other Josephson variable, and (2) it shares an extraneous measurement component with the Whyte measure.

17 In this respect when constructed scales are used to test hypotheses, it is especially important to inspect correlations and assess scale reliability when the items do not form a single factor – see Nunnally (1967: 237).

It is also important to consider the source of systematic errors or threats to the validity of measurement. Nunnally (1967: 206-210) discusses such threats in the testing context, distinguishing between variation within a test (e.g., different sampling of items, guessing, situational factors, and clerical or scoring errors, all of which can be encompassed within the domain-sampling model), and between tests (differences in content, subjectivity of scoring, and actual change, which cannot be handled adequately by a reliability model based on random sampling of items). In the latter case, however, reliability theory can be extended to sampling whole tests; "and correlations among tests are permitted to be somewhat lower than predicted from the correlations among items within tests. In this case the average correlation among a number of alternative forms administered on different occasions. . . would be a better estimate of reliability" (Nunnally 1967: 210).

18 Note that reliability analysis cannot be done on the reliability-by-case subsample, since the subsample is not selected independently of the reliability analysis.

REFERENCES

Barry, Herbert, ill, L. Josephson, E. Lauer, and C. Marshall 1976 Traits Inculcated in Childhood: Cross-Cultural Codes 5. Ethnology 15: 83-114.

Barry, Herbert, ill, L. Josephson, E. Lauer, and C. Marshall 1977 Agents and Techniques for Child Training: Cross-Cultural Codes 6. Ethnology 16: 191-230.

Barry, Herbert, ill and Leonora M. Paxson 1971 Infancy and Early Childhood: Cross-Cultural Codes 2. Ethnology 10: 466-508.

Barry, Herbert, ill and Alice Schlegel 1982 Cross-Cultural Codes on Contributions by Women to Subsistence. Ethnology 21: 165-188.

Batchelder, William H. and A. K. Romney 1988 Test Theory without an Answer Key. Psychometrika 53(1): 71-92.

Batchelder, William H. and A. K. Romney 1989 New Results in Test Theory without an Answer Key. Mathematical Psychology in Progress, ed. E. E. Roskam, pp. 229-248. Springer-Verlag.

Bernard, H. Russell 1988 Research Methods in Cultural Anthropology. Newbury Park: Sage.

Bradley, Candice 1989 Reliability and Inference in the Cross-Cultural Coding Process. Journal of Quantitative Anthropology 1: 353-371.

Brown, W. 1910 Some Experimental Results in the Correlation of Mental Abilities. British Journal of Psychology 3: 296-322.

Campbell, Donald T. 1970 Natural Selection as an Epistemological Model. In R. Naroll and R. Cohen, eds., Handbook of Cultural Anthropology, pp. 51-85.

Campbell, Donald T. and Donald W. Fiske 1959 Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychological Bulletin 56: 81-105.

Comrey, Andrew L. 1962 The Minimum Residual Method of Factor Analysis. Psychological Reports 11: 15-18.

Fisher, R. A. 1948 Statistical Methods for Research Workers. New York: Hafner Publishing Company.

Harman, Harry H. 1967 Modern Factor Analysis. Chicago: University of Chicago Press.

Hote1ling, H. 1933 Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology 24: 417-520.

Jester, Ralph B. 1989 Family Structure and the Belief in High Gods. Ms. University of California, Irvine.

Johnson, Allen and Clifford Behrens 1989 Time Allocation Research and Aspects of Method in Cross-Cultural Comparison. Journal of Quantitative Anthropology 1: 313-334.

Kelley, T. L. 1924 Statistical Methods. New York: MacMillan. Kim, Jae-On and Charles W. Mueller 1978 Factor Analysis: Statistical Methods and Practical Issues. Beverly Hills: Sage.

Kirk, Jerome and Marc Miller 1986 Reliability and Validity in Qualitative Research. Beverly Hills.

Lord, Frederic M. 1980 Applications of Item Response Theory to Practical Testing Problems. Hillside, NJ: Lawrence Erlbaum.

Lord, Frederic M. and Melvin R. Novick 1968 Statistical Theories of Mental Test Scores. Reading, MA: Addison Wesley.

Murdock, George P. and 1969 Standard Cross-Cultural Sample. Ethnology 8: 329-369.

Naroll, Raoul 1962 Data Quality Control. New York: The Free Press.

Nunnally, J. C. 1967 (2nd edition 1978). Psychometric Theory. New York: McGraw-Hill.

Pearson, K. 1896 Mathematical Contributions to the Theory of Evolution, Ill: Regression, Heredity and Panmixia. Philosophical Transactions of the Royal Society of London (A) 187: 253-318.

Pelto, Pertti J. and Gretel H. Pelto 1978 Anthropological Research: The Structure of Inquiry. Cambridge: Cambridge University Press.

Rohner, Ronald P. 1981 They Love Me, They Love Me Not. New Haven: HRAF Press.

Rohner, Ronald P., Scott D. Berg, and Evelyn C. Rohner 1982 Data Quality Control in the Standard Cross-Cultural Sample: Cross-Cultural Codes. Ethnology 21: 359-372.

Rohner, Ronald P. and Evelyn C. Rohner 1982 Enculturative Continuity and the Importance of Caretakers: Cross-Cultural Codes. Behavior Science Research 17: 91-114.

Romney, A. K. 1989 Quantitative Models, Science and Cumulative Knowledge. Journal of Anthropological Anthropology 1: 153-223.

Romney, A. K., William H. Batchelder, and Susan C. Weller 1987 Recent Applications of Cultural Consensus Theory. American Behavioral Scientist 31(2): 163-177.

Romney, A. K. and Susan C. Weller 1989 Systematic Culture Patterns in High Concordance Codes. Roberts Feschrift, R. Bolton, ed. HRAF Press.

Romney, A. K., Susan C. Weller, and William H. Batchelder 1986 Culture as Consensus: A Theory of Culture and Informant Accuracy. American Anthropologist 88: 313-338.

Schuessler, Karl 1971 Analysing Social Data: A Statistical Orientation. Boston: Houghton Mifflin Co.

Spearman, C. 1904a The Proof and Measurement of Association between Two Things. American Journal of Psychology 15: 72-101.

Spearman, C. 1904b 'General Intelligence,' Objectively Determined and Measured. American Journal of Psychology 15: 201-293.

Spearman, C. 1910 Correlation Calculated from Faulty Data. British Journal of Psychology 3: 271-295. Spearman, C. 1927 The Abilities of Man. New York: MacMillan.

Spiro, Melford E. 1986 Cultural Relativism and the Future of Anthropology. Cultural Anthropology 1(3): 259-286.

Terry, Roger L. 1971 Dependence Nurturance and Monotheism: a Cross-Cultural Study. Journal of Social Psychology 84: 175-181.

Toulmin, Stephen 1981 Evolution, Adaptation and Human Understanding. In M. B. Brewer and B. E. Collins, eds., Scientific Inquiry and the Social Sciences. San Francisco: Jossey_Bass.

Tukey, John 1977 Exploratory Data Analysis. Reading, Mass.: Addison-Wesley.

Tryon, Robert C. 1957 Reliability and Behavior Domain Validity: Reformulation and Historical Critique. Psychological Bulletin 54: 229-249.

Weller, Susan C. and Romney, A. K. 1988 Systematic Data Collection. Sage Publications.

Weller, Susan C. and Romney, A. K. n.d. Metric Scaling: Correspondence Analysis.

White, Douglas R. 1990 CORR-REL: A Program for Reliability Analysis and Optimal One-Factor Scaling. World Cultures, forthcoming.

Whiting, John M. W. 1981 Environmental Constraints on Infant Care Practices. R. Munroe, L. Munroe, and B. Whiting, eds., Handbook of Cross-Cultural Human Development, pp. 155-179.

Whyte, Martin K. 1978 Cross-Cultural Codes Dealing with the Relative Status of Women. Ethnology 17: 211-37.

Yule, G. U. 1897 On the Theory of Correlation. Journal of the Royal Statistical Society 60: 812-854.

Yule, G. U. 1899 An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades. Journal of the Royal Statistical Society 20: 249-295.

Yule, G. U. 1922 An Introduction to the Theory of Statistics. London: Griffin.

Page 116 line 2 : F hat → $\hat{F}$ Page 117 line 2 : F hat → $\hat{F}$
FOOTNOTE PLACEMENTS
113: 1-2 116: 3-6 118: 7 119: 8-9 120: 10 121: 11 122 :12 123: 13
127: 14 129: 15 140: 16 141: 17 143: 18

There are probably still lots of notational misprints in this version that were not in the published version.