Reliability and Measurement Error in the Presence of Homogeneity

Cathy King Pike and Walter W. Hudson

Abstract

This paper describes a limitation of using Cronbach's Alpha to estimate reliability when using a sample with homogeneous responses in the measured construct. More specifically, it describes the risk of falsely concluding that a new instrument may have poor reliability and demonstrates the use of an alternate statistic that may serve as a cushion against such errors. Data from two validation studies are used to illustrate the utility of the new statistic, referred to as R-Alpha or Relative Alpha**.** Included is a discussion of the limitations and appropriate use of the statistic in validating multi-item tests, assessment scales, and inventories.

Although there are many different measures of reliability, Cronbach's (1951) coefficient alpha has come to be a standard and preferred means of computing a reliability coefficient for many types of multi-item unidimensional and multidimensional assessment scales, tests, and inventories. The reason for this strong preference arises from the fact that the coefficient alpha has a number of highly desirable characteristics. Alpha is easy to compute from the standard deviations for items and the standard deviation for the total score on a measurement scale. Alpha is the mean of all possible split-half reliability coefficients and can be used to estimate parallel test reliability. Alpha also serves as a confirmation of unidimensionality when its value for a given instrument exceeds .90. Moreover, Nunnally (1978) illustrated how Cronbach's alpha can be directly derived from the older Generalized Spearman-Brown (GSB) formula and that they are identical except for a change in metric (2-scores for GSB and raw scores for Alpha).

Despite the great simplicity, power, and utility of the coefficient alpha, it can present problems for applied researchers who seek to develop and validate new assessment tools using the domain sampling model of classical measurement theory. A pronounced feature of Alpha is that it is based on inter-item correlations through its derivation from the GSB formula, so its value can, and often is, affected by homogeneity of subject responses to scale items. The effect of pronounced homogeneity is to attenuate the magnitude of inter-item correlations and, hence, one's estimate of alpha based on the sample at hand.

For example, in the initial validation research for the Multi-Problem Screening Inventory or MPSI scale (Hudson & McMurphy, 1990; Hudson, 1990), we investigated the value of the coefficient alphas for all 27 of the subscales of the MPSI. The aim in designing the MPSI scale was to have alpha equal or exceed .80 for all 27 of the MPSI subscales, and that was clearly achieved for 26 of them. In fact, 16 of the subscales had coefficients of .90 or better. Unfortunately, an important subscale purporting to measure aggression produced a disappointing alpha of only .71, and it was feared the Aggression subscale of the MPSI was flawed. Investigation of this dilemma took note of the fact that the sample of more than 300 respondents were social work students, most of whom were studying for a master's degree. It then occurred to us that, as a population, social workers are clearly one of the least aggressive self-selected populations one might identify, along with perhaps nurses and ministers.

As an alternative to evaluating the reliability of the Aggression subscale for the MPSI, we computed the standard error of measurement as follows: SEM = s * Sqrt (1 -Alpha), where “s” is the standard deviation of the Aggression subscale. In doing so, we found that the Aggression subscale of the MPSI had an SEM =2.70. Since the SEM is an estimate of the standard deviation of the errors of measurement and it is in the raw score metric of the subscale, we were very encouraged by the small SEM for the Aggression subscale, given the fact that scores could range from 0 to 100. Although the Cronbach's coefficient alpha for this subscale was disappointingly small, the SEM strongly suggested that the subscale had very good measurement error characteristics. That is, it appeared to us that the low value of alpha was not caused by measurement deficiencies, but by the marked homogeneity of the sample with respect to the attribute in question-aggression. We therefore decided to retain the Aggression subscale without modification, and we did so with the belief that subsequent testing with a more heterogeneous sample would produce a much larger value of alpha. Indeed, a recent study of veterans yielded a coefficient alpha for the Aggression subscale of .88 (Pike & 'Hudson, in press).

The point of this example is to caution against a premature conclusion that a new measurement tool is unworthy merely because it has a low coefficient alpha. At a minimum, one should always compute the values of alpha and the SEM. If one does find a very low alpha and a very high SEM, the best one can do is recognize that the instrument in question does not have good measurement error characteristics. On the other hand, if one has a fairly low alpha but also a relatively low SEM, that may be one signal that the instrument does have good or at least acceptable measurement error characteristics. The low coefficient alpha could have resulted from marked homogeneity in the item responses by those in the sample (not the population) with respect to the measured characteristic.

Estimating the coefficient alpha

Nunnally (1978) proposed a means of estimating the coefficient alpha in the case of sample homogeneity. The formula for estimating alpha is the same formula used in calculating any alpha, Alpha = 1 – SEM^2/S_0^2**,** except that a hypothesized variance representing a more heterogeneous sample is substituted for the total score variance obtained from the homogeneous sample (Nunnally & Bemstein, 1994, p. 261). To obtain this estimate one simply divides the error variance by the hypothesized variance and subtracts the quotient from 1 .0. However, one must be able to reasonably estimate the variance that might be expected from a more heterogeneous sample to use this formula in estimating alpha. The accuracy of this estimate for a more heterogeneous sample is dependent upon the usually sensible assumption of relatively equal standard errors of measurement (Nunnally & Bernstein, 1994). However, if a single, large homogeneous sample was used, it may be impossible to reasonably estimate the variance of a more heterogeneous sample.

We propose a means of estimating alpha, using the SEM and range of a given scale. The SEM reflects the amount of measurement error in a scale and is less dependent upon the degree of homogeneity. However, the SEM when used alone is limited as a means of estimating the coefficient alpha. This is because the value of the SEM is unique to the metric of a particular scale and must be interpreted within the possible range of scale scores. Alpha will always have a range from 0 to 1, but the SEM will have a range that is unique to its instrument. At stake here is the hope of obtaining an interpretation of the SEM that will give us a better sense of how much error is present in a measurement scale relative to some benchmark. A good starting point is to recognize that the SEM must be small, in some sense, in order to assert that acceptable measurement error characteristics have been achieved. We suggest that the SEM should be small in relation to the total range of scores that might be achieved with a particular measurement device. Stated differently, if the SEM is a small proportion of the possible range of scores that can be obtained from the measurement tool, it would seem reasonable to believe that the presence of measurement error is not a serious handicap of the instrument.

Truncation of Range and Variance

Loether and McTavish (1988) stated that the total score variance, used in calculating the coefficient alpha with the previous equation, “can be interpreted in terms of a scale extending from a minimum possible value that equals the range divided by the square root of twice the sample size” (p. 151). They report that “for most curves, the range is approximately equal to six times the standard deviation as a rough rule of thumb” (p. 151) and we note that this holds only for normally distributed values. If we follow this reasoning, there is merit in more closely examining the formula for alpha, shown previously, with special attention given to the rightmost term in the equation.

Here we see that measurement error is expressed as a simple proportion of the total score variance. From this vantage point it can be seen that a small total score variance arising from use of a homogeneous sample of total score responses will produce a large fraction and, hence, a small alpha coefficient. This points directly to a very important question, “Does the observed total score variance, S_0^2, do a good job of representing the expected total score variance, S_e^2, over the entire range of possible scores?” In this regard, we can turn back to Loether and McTavish (1988) and recognize that the expected total score standard deviation for a set of normally distributed scores can be estimated as S_e = Range/6 and the expected range can be estimated as $\text{Range}_e = 6 * S_{0.}$ This means that we can examine the approximate truncation in range from use of a homogeneous sample by computing the truncated range, TR, as TR = $\text{Range}_e/\ \text{Range}_0$ where the observed range is simply the highest possible score minus the lowest possible score. In the case of the MPSI Aggression subscale we can compute the value of $S_0{}^2$ from the relation, S_0 = SEM/Sqrt(1 -alpha)

= 2.70/Sqrt(l - .71)

= 5.01

so the expected range is computed as Range, = 5.01 * 6

= 30.06

and TR = $\text{Range}_e/\text{Range}_0$

= 30.06/60
= 5010

In other words, there has arisen an approximate 50% shrinkage in the effective range of the MPSI Aggression subscale as a consequence of having obtained a homogeneous sample. Stated differently, the computed value of TR provides direct evidence of shrinkage in coefficient alpha due to the use **of** a homogeneous sample of scale scores.

Range computations

Before continuing, it is important to note that the MPSI subscales have scores that range from 0 to 100 but coefficient alpha is always computed using simple sum scores. For example, the MPSI Aggression subscale contains 10 items and each is scored over a range from ‘1’ to **‘7’.** This means that the raw sum score for this subscale ranges from a low of 10 to a high of 70, so the actual range of the sum scores is 60 = 70 - 10. When using the MPSI with clients, the sum scores are transformed, for ease of interpretation, to range from 0 to 100 rather than 10 to 70. For purposes of this analysis we shall always compute score ranges based on raw sum scores rather than transformed sum scores. Thus, the lowest and highest possible sum score for any scale, and hence the range, will depend on the number of items in each scale and the range of values over which each item is scaled.

Coefficient R-Alpha

The foregoing does little more than provide an analysis of coefficient alpha with a view toward better understanding how a sample of homogeneous scale scores can truncate the effective range of scores which, in turn, attenuates inter-item covariances and correlations and, hence, the magnitude of alpha. In seeking some protection against the risk of falsely concluding that a new scale has unacceptable measurement error characteristics because it has an unacceptable alpha, we turn again to Loether and McTavish (1988) and capitalize on Eq. 2 which combines with the formula for alpha in Eq. 1 to produce a statistic we call “R-Alpha” where R-Alpha = 1 – SEM^2/(Range/6)^2 We call it the “Relative Alpha” because it is an estimate of the obtained error variance relative to what we would expect for a set of normally distributed scores randomly sampled over the possible scores for the scale. Recall that Nunnally and Bernstein (1994) suggested that we compute alpha using some hypothesized total score variance. As we noted earlier, one must be able to reasonably estimate the total score variance that might be expected from a more heterogeneous sample in order to use this formula in estimating alpha; we do that by merely dividing the possible range of scores by 6 and squaring the result. This gives us an estimate of measurement error in relation to possible scores that can be achieved on an instrument. In the above example, we obtained an SEM of 2.70 for the Aggression subscale of the MPSI. Since the Aggression subscale is scored over a range from 10 to 70, the range equals 60 and the obtained SEM^2, relative to the expected total score variance, yields a value of 2.702/(60/6)2 = 0.0729. That is gratifyingly small proportion of expected total score error variance, but what we really seek is an estimate of measurement error that reflects non-error or "reliability." We, therefore, compute R-Alpha $= 1 – \text{SEM}^2/(\text{Range}/6)^2$

$= 1 - 2.70^2/(60/6)^2$

$= 1 - ,0729$

$= 0.9271$

The mathematical justification for this formula lies in the simple fact that scores cannot fall beyond the minimum and maximum possible values (the range) on a closed interval measure such as is produced by summated category partition scales (Stevens, 1968) or so-called Likert scales, However, it should be noted that the value of TR can sometimes exceed 1.0, because the estimated range as computed from R, $= S_0 * 6$ will exceed the possible range of scores for the particular scale. When that occurs, the value of R-Alpha will be smaller than the observed alpha. Either event (R-Alpha < alpha and TR > 1.0 **)** strongly suggests that a small observed alpha is not due to a problem with homogeneous responses. Three such instances can be seen in Table 2. Most important, R-Alpha is an estimate of what alpha is likely to be if one collects a fresh sample of data and insures that the new scores have an observed total score standard deviation approximately equal to Range/6.

An extended example

In the example of the Aggression subscale of the MPSI, we have the benefit of our rationale, our findings, and the confirmation from a fresh sample (believed, a priori, to be more heterogeneous) which showed that the subscale did achieve an acceptable reliability of alpha = .88 (Pike & Hudson, in press). In further investigation of R-Alpha and its possible utility for applied validation research, we also have additional data for a new measurement device that provides a second opportunity to examine reliability and measurement error evaluations in the face of homogeneous responses to the measured construct.

In this study, Pike (1994) developed the Social Work Values Inventory (SWVI). This instrument was developed to measure adherence to the four most commonly cited professional values. The four values were confidentiality, self-determination, dignity and worth, and social justice. Preliminary testing revealed that the Dignity and Worth Scale loaded across the other three scales. Because dignity and worth seemed to underlie the three remaining values, it may represent a value orientation rather than a value. This scale subsequently was dropped from the SWVI, in order to maintain a focus on values, and will not be reported in this paper.

The items of the SWVI were presented as practice vignettes in which one of the four values is called into question. Using a five point scale, a graduated continuum of extreme positions on the value in question was presented. Respondents indicated the degree to which the social worker in the vignette should be oriented toward one extreme position or the other.

Four pilot tests of the SWVI were conducted where three of these examined the internal consistency reliability of the scales (Pike, 1994). The first pilot test was completed with a small sample (n = 24) of baccalaureate and master's level students of two southern universities. The sample for the second pilot test (n = 31) was comprised of field instructors holding the MSW degree and having at least two years of social work experience. The data for the third pilot test was collected using a mail survey of 400 NASW members.

Estimates of internal consistency were computed using the items of the scales as conceptualized and then deleting the item that would yield the highest overall alpha if deleted. This process was continued until no item in the scale would yield a higher scale coefficient alpha, if deleted. Thus, the number of items remaining in each scale differed across the three samples, depending upon their utility as indicators of the construct. Studies of SWVI that have examined its validity have found superb evidence of content, factorial, discriminant and known groups construct validity (Pike, 1996; Rice, 1994).

Table 1 lists the descriptive statistics for the three studies by Pike (1994). Table 2 contains the SEMs, coefficient alphas, R-Alphas, and coefficients of variation for each of the scales. An examination of the coefficient alphas across the pilot tests suggests a loss to the level of alpha with each of the two subsequent pilot tests. In contrast, the SEMs remain relatively low and stable across the three pilot tests. Also noteworthy is the substantial drop in the standard deviations of the last .two pilot tests. To examine the extent of variation across the pilot test, a normed measure of variation was computed (Martin & Gray, 1971). This index provides a relative measure of variation across samples by dividing the standard deviation by the mean, and then dividing that quotient by the square root of N - 1. The index can range from a minimum of 0 to a maximum of 1.0. An examination of the coefficients of variation across the pilot tests indicates that the relative variation across the three samples is extremely low and decreases still further in the last two samples.

A review of Table 2 suggests that all of the scales of the SWVI have good measurement error characteristics, yielding estimates of the coefficient alphas in the good to excellent range. The R-Alpha in most cases reasonably approximates the actual coefficient obtained in the first pilot test. However, it should be remembered that the first pilot test contained responses from a very small sample, and sample size could have served to restrict the variation in that sample. Any restriction of variation due to the small sample size would serve to attenuate the observed coefficient alphas.

If only the coefficient alphas had been examined in the SWVI pilot tests, the SWVI would have seemed to have had unacceptable measurement error characteristics. Upon further examination, and finding results indicative of good measurement error characteristics, the question was asked: “What, besides measurement error, could have resulted in a substantial drop in the coefficient alphas across these three samples?” If one assumes that professionals of a discipline share common values of that profession, data collection using an instrument developed to measure those values reasonably would result in restricted ranges and attenuated coefficient alphas. The coefficients of variation that were calculated for each sample showed that sample responses became even more homogeneous as educational and experience levels in social work increased. This homogeneity likely resulted in reductions to the coefficient alphas. The first sample of respondents were baccalaureate and master’s students majoring in social work while the last two samples were entirely comprised of MSW practitioners. All respondents in the field instructor sample held the MSW degree and had at least two years of social work experience. Some, but not all, of the field instructors were members of NASW**.** The NASW membership sample consisted of highly experienced practitioners. Of the 195 participants, 181 held the MSW degree, while five others held doctorates in social work or another area. In addition to the differences in educational level and extent of social work experience, self-selection through NASW membership also may have contributed homogeneity in responses (Gibehan & Shervish, 1993; Judd, Block, & Jain, 1985), in that NASW is a politically liberal organization and does not represent all social workers in the US.

The finding that further education and experience increases the extent to which social workers agree about the enactment of professional values is crucial knowledge in specifying the role of values in professional socialization. Further, the SWVI is the first instrument measuring professional values in social work that has been capable of distinguishing significant differences across levels of education (Rice, 1994). A decision, based on the reductions to the coefficient alphas, to cease further research of the SWVI would have prevented inquiry in an area of substantial importance to social work education.

Conclusions

As can be seen from the foregoing examples, the R-Alpha statistic may be very useful in indicating that a measurement tool does have acceptable measurement error characteristics in the face of low or even discouraging coefficient alphas. However, we hasten to note that the R-Alpha statistic should never be considered as a substitute for Cronbach's alpha. We suggest R-Alpha merely as a temporary device that may be of use when it is strongly suspected that low coefficient alphas have occurred because of marked homogeneity of responses to items in the sample at hand. In short, R-Alpha can never be the final arbiter in making decisions about instrument reliability. Ultimately, that must be done with a fresh sample of data in which an effort has been made to obtain the heterogeneity of item responses that will avoid unreasonable attenuation of inter-item correlations.

We clearly acknowledge that R-Alpha is not a replacement for alpha. In fact, R-Alpha shown in Eq. 6 is alpha except for the use of $(\text{Range}/6)^2$ in the denominator of Eq. 1 instead of the usual value for S_0^2. However, the R-Alpha's assumption of sample distribution may be incorrect for a given sample and, in this case, can provide misleading results when only one sample is available for examination.

Despite the need for careful use of R-Alpha to avoid making overblown claims for reliability, we believe it has much to offer. First, it enables one to capitalize on small-sample studies when they have the limited purpose of exploring the feasibility of conducting a full-scale psychometric evaluation. Although small-sample pilot tests lack parameter estimation power, they are considered very useful in the early development of an instrument (Converse & Presser, 1986; Rossi, Wright & Anderson, 1983). Their speed, efficiency, and low cost allow the researcher to examine the potential merits of a new instrument while conserving resources for final psychometric testing.

The second major benefit of R-Alpha is the one that motivated this article. It can be a powerful protection against falsely concluding that a new measurement tool has poor measurement error characteristics because computed alphas have artificially low values due to response homogeneity. Prematurely discarding a new measurement tool has conspicuous costs and consequences, and R-Alpha can help avoid these.

Finally, R-Alpha is a very useful tool for examining the impact on reliability of increasing the number of response categories for the items in a scale. Before conducting further research using the SWVI, R-Alpha was used to estimate the effect on the coefficient alphas when the response categories were increased from 5 to 7 points. A subsequent pilot test of the 7-point response categories resulted in higher coefficient alphas, even though only a small sample (N = 37) had been available to participate in the pilot study. In this study, the Confidentiality, Self-Determination, and Social Justice Scales yielded the following coefficient alphas, respectively: .72; .78; and, .87. Further research on the SWVI with a large sample is planned and should provide more definitive information about the internal consistency of the SWVI Scales using the 7-point response categories.

We recommend that researchers use the Nunnally formula by estimating the variance for a heterogeneous sample as $(\text{Rar1ge}/6)^2$ and that is all that we have done in this paper; we refer to this estimate as R-Alpha to denote the very specific way in which the variance estimate is made. Ultimately, researchers must resolve the reliability question with a fresh sample of data and with new estimates of Cronbach’s alpha. Used with due regard to the inherent limitations to estimation, these procedures may prevent the mistake of discarding potentially useful measurement tools.

Author note

Because we expect to use R-Alpha in our own work, we have developed an R-Alpha program for use with Windows 3.11, Windows NT, or Win95. The program is available without fee and can be downloaded from the World Wide Web using the URL, http://www.indirect.com/www/walmyr/wpchome.htm **.** you do not have access to the Web, you can obtain the R-Alpha software for the cost of materials, shipping, and handling. Send a check in the amount of $7.00 to the WALMYR Publishing Co. (P.0. Box 6229, Tallahassee, FI 32314), and they will send you a copy of the R-Alpha program.

List of tables

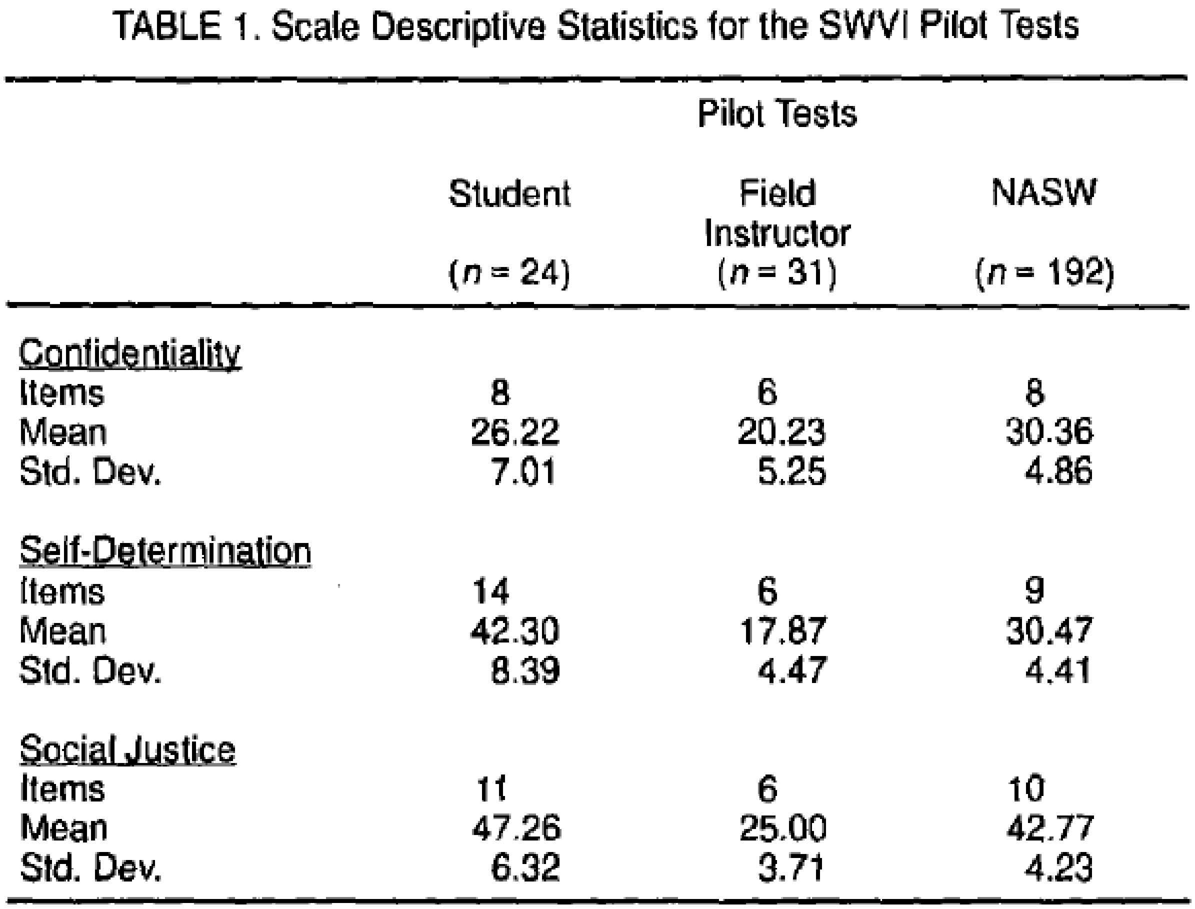

TABLE 1. Scale Descriptive Statistics for the SWVI Pilot Tests

	Pilot Tests		
	Student (*n* = 24)	Field Instructor (*n* = 31)	NASW (*n* = 192)
Confidentiality			
Items	8	6	8
Mean	26.22	20.23	30.36
Std. Dev.	7.01	5.25	4.86
Self-Determination			
Items	14	6	9
Mean	42.30	17.87	30.47
Std. Dev.	8.39	4.47	4.41
Social Justice			
Items	11	6	10
Mean	47.26	25.00	42.77
Std. Dev.	6.32	3.71	4.23

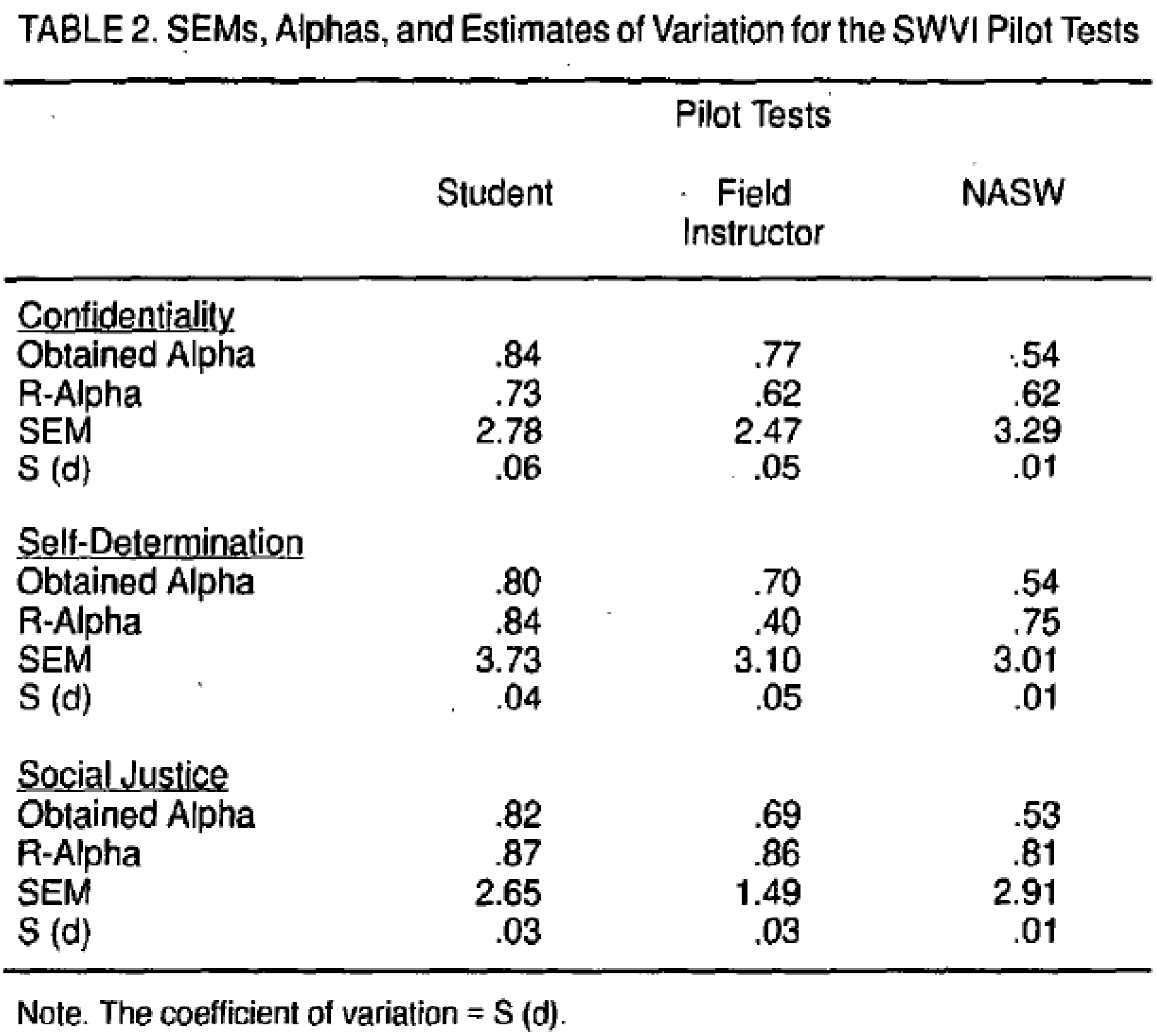

TABLE 2. SEMs, Alphas, and Estimates of Variation for the SWVI Pilot Tests

	Pilot Tests		
	Student	Field Instructor	NASW
Confidentiality			
Obtained Alpha	.84	.77	.54
R-Alpha	.73	.62	.62
SEM	2.78	2.47	3.29
S (d)	.06	.05	.01
Self-Determination			
Obtained Alpha	.80	.70	.54
R-Alpha	.84	.40	.75
SEM	3.73	3.10	3.01
S (d)	.04	.05	.01
Social Justice			
Obtained Alpha	.82	.69	.53
R-Alpha	.87	.86	.81
SEM	2.65	1.49	2.91
S (d)	.03	.03	.01

Note. The coefficient of variation = S (d).

References

1. Converse, J. M. & Presser, S. (1986). Survey Questions: Handcrafting the Standardized Questionnaire. Quantitative Applications in the Social Sciences, Vol. 63. Newbury Park, CA: Sage.
2. Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika, 6, 291-334.
3. Gibelman, M. & Shervish, P. H. (1993). Who We Are: The Social Work Labor Force as Reflected in the NASW Membership. Washington, DC: National Association of Social Workers.
4. Hudson, W. W. (1990). The MPSI Technical Manual. Tempe, AZ: Walmyr Publishing Co.
5. Hudson, W. W. & McMurtry, S. L. (1997). Comprehensive Assessment in Social Work Practice: The Multi-Problem Screening Inventory. Research on Social Work Practice, 7, 79-98.
6. Judd, P., Block, S. R., & Jain, A. K. (1985). Who Joins NASW: A Study of Graduating MSWs. Arete, 10(2), 41-44.
7. Loether, H. J. & McTavish, D. G. (1988). Descriptive and Inferential Statistics: An Introduction (3rd ed.). Boston: Allyn and Bacon.
8. Martin, J. D. & Gray, L. N. (1971). Measurement of Relative Variation: Sociological Examples. American Sociological Review, 36,496-502.
9. Murphy, D. L. (personal communication, 1994).
10. Nunnally, J. C. (1978). Psychometric Theory (2nd ed.). New York: McGraw-Hill.
11. Nunnally, I. C. & Bernstein, I. H. (1994). Psychometric Theory (3rd ed.). New York: McGraw-Hill.
12. Pike, C. K. (1994). Development of the Social Work Values Inventory (Doctoral dissertation, University of Alabama, 1994). Dissertation Abstracts International.
13. Pike, C. K. (1996). Development and Initial Validation of the Social Work Values Inventory. Research on Social Work Practice, 6,337-352.
14. Pike, C. K. (in press). Using Second-order Factor Analysis in Examining the Multiple Problems of Clients. Research on Social Work Practice.
15. Rice, D. S. (1994). Professional Values and Moral Development: The Social Work Student (Doctoral dissertation, University of South Carolina, 1994). Dissertation Abstracts International.
16. Rossi, P. H., Wright, J. D., & Anderson, A. B. (Eds.) (1983). Handbook of Survey Research. San Diego: Academic Press.
17. Smith, P. K. & Kendall, L. M. (1963). Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales. Journal of Applied Psychology, 47, 149-1 55.
18. Stevens, S. S. (1968). Ratio Scales of Option. In D. K. Whitla (Ed.), Handbook of Measurement and Assessment in Behavioral Science, 171 - 199. Reading, MA: Addison-Wesley.