2002

Reliability generalization of Working Alliance Inventory scale scores

W. E Hanson
K. T Curry
Deborah L Bandalos, *James Madison University*

RELIABILITY GENERALIZATION OF WORKING ALLIANCE INVENTORY SCALE SCORES

WILLIAM E. HANSON, KYLE T. CURRY,
AND DEBORAH L. BANDALOS
University of Nebraska–Lincoln

Reliability generalization (RG) was used to study five versions of the Working Alliance Inventory (WAI), including scores from 12 different scales. Sixty-seven internal consistency estimates, six interrater reliability estimates, and four study characteristics were analyzed. In general, reliability estimates of WAI scale scores appear to be robust. Mean reliability estimates ranged, in this sample of studies, from .79 to .97, with a modal estimate of .92. Variability in reliability estimates was, based on simple bivariate correlations, associated with client and therapist sample size for WAI total scores (observer version). Implications for measuring alliance using the WAI and conducting future RG studies on psychotherapy process measures are discussed.

The outcome question, “Is psychotherapy effective?” has been answered. At least five decades of research suggest that the answer is yes, it is effective (Lambert & Bergin, 1994; Smith, Glass, & Miller, 1980; Wampold, 2001). However, the process question, “What makes psychotherapy effective?” has not been answered, at least not satisfactorily. Consequently, the need for research on variables that affect the psychotherapy change process (e.g., client and therapist variables, common factors, etc.) remains strong.

One variable in particular, the so-called working alliance (WA), has received considerable attention in the psychotherapy literature recently and has emerged as a promising construct for studying and better understanding the change process (Gaston, 1990). The WA, defined here as the extent to which a client and therapist work collaboratively and purposefully and con-

The authors gratefully acknowledge Patricia Cerda and Chris Gibbon, who helped identify, locate, and code published WAI studies. Correspondence regarding this article should be addressed to William E. Hanson, Department of Educational Psychology, 228 TEAC, University of Nebraska–Lincoln, Lincoln, NE 68588-0345; e-mail: whanson1@unl.edu.

Educational and Psychological Measurement, Vol. 62 No. 4, August 2002 659-673

nect emotionally, is conceptualized as a common, or generic, factor in that it is believed to cut across various treatment approaches (Bordin, 1979; Greenson, 1965; Horvath & Greenberg, 1994; Horvath & Luborsky, 1993; Wolfe & Goldfried, 1988). Wolfe and Goldfried (1988), for example, described it as the “quintessential integrative variable” (p. 449). In dozens of studies, WA has been found to be correlated positively with a broad range of psychotherapy outcomes and, overall, appears to be a relatively strong predictor of client change (for reviews, see Horvath & Symonds, 1991; Martin, Garske, & Davis, 2000).

Since the early 1980s, a number of instruments have been developed to measure WA. Currently, at least seven such instruments exist in the public domain. These instruments, listed alphabetically, are the California Psychotherapy Alliance Scales (CALPAS) (Marmar, Weiss, & Gaston, 1989), the Penn Helping Alliance Rating System (HA_r) (Luborsky, Crits-Christoph, Alexander, Margolis, & Cohen, 1983), the Therapeutic Alliance Scale (TAS) (Marziali, 1984), the Therapeutic Bond Scales (TBS) (Saunders, Howard, & Orlinsky, 1989), the Vanderbilt Therapeutic Alliance Scale (VTAS) (Hartley & Strupp, 1983), the Working Alliance Inventory (WAI) (Horvath, 1981; Horvath & Greenberg, 1986, 1989), and the Working Alliance Inventory–Short (WAI-S) (Tracey & Kokotovic, 1989). The primary purpose of this study was to examine the reliability generalization (RG) of the scale scores of two of these instruments: the WAI and the WAI-S.

The WAI was one of the first instruments of its kind—certainly one of the first to attempt to quantify the various theoretical dimensions of WA (Horvath & Greenberg, 1986). It was selected for several reasons. First of all, it is by far the most popular measure of WA available. Second, it is a self-report instrument that can be administered easily and completed rapidly, either by a client, a therapist, or a nonparticipant observer. Third, as noted, it is theoretically based. Fourth, its scale scores have been shown to share a significant amount of common variance with other measures of WA (Tichenor & Hill, 1989). And finally, it is familiar to psychotherapy process researchers, as well as to clinicians. The WAI-S was selected because it aligns closely with its parent instrument, the WAI.

Before describing these two instruments and the known psychometric properties of each of their scale scores, RG will be discussed briefly. This brief discussion is intended to give context to the study and to help the reader better understand the methods used.

RG

Given that test scores, not tests, vary in their reliability, it is important to study systematically the reliability estimates of study-specific test scores (Crocker & Algina, 1986; Dawis, 1987; Feldt & Brennan, 1989; Pedhazur &

Schmelkin, 1991; Rowley, 1976; Thompson, 1994; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998; Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999). To this end, Vacha-Haase (1998) introduced a meta-analytic method that she called RG. According to Vacha-Haase, RG can be used to "characterize the mean measurement error variance across studies, and also the sources of variability of these variances across studies" (p. 6). In other words, it can be used to evaluate the robustness of a given test's score reliability, as well as factors that may affect the reliability of a test's scores.

To conduct an RG study, reliability estimates and study characteristics have to be accumulated, coded, and analyzed statistically. The problem, however, is that this type of information is not always reported in published studies (see Meier & Davis, 1990; Vacha-Haase, Kogan, & Thompson, 2000; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). Instead, researchers tend to rely on reliability induction, whereby test score reliability is generalized inductively, usually from a single sample, and assumed to be an appropriate estimate for other samples (Vacha-Haase et al., 2000). Moreover, study characteristics are, at best, reported incompletely or, at worst, reported not at all. Unfortunately, this handcuffs would-be RG researchers, making it difficult for them to locate enough usable studies to conduct a meta-analysis of this sort.

Nevertheless, RG studies have been conducted on a number of well-known, clinically-oriented psychological tests, including, to name a few, symptom-based tests, such as the Beck Depression Inventory (BDI) (Yin & Fan, 2000), and personality tests, such as the Minnesota Multiphasic Personality Inventory (MMPI/MMPI-2) (Vacha-Haase, Kogan, Tani, & Woodall, 2001) and the NEO Personality Inventory–Revised (NEO-PI-R) (Caruso, 2000). Somewhat surprisingly, no RG studies have been conducted on any of the existing psychotherapy change process measures, such as the WAI or the WAI-S. These measures are described in the next two sections.

WAI

The WAI is a 36-item self-report measure of WA (Horvath, 1981; Horvath & Greenberg, 1986, 1989). It has three subscales: Goals, Tasks, and Bond, each of which is based on Bordin's (1979) multidimensional theoretical conceptualization of WA. The Goals subscale measures the extent to which a client and therapist agree on the "goals (outcomes) that are the target of the intervention" (Horvath & Greenberg, 1989, p. 224). The Tasks subscale measures the extent to which a client and therapist agree on the "in-counseling behaviors and cognitions that form the substance of the counseling process" (p. 224). The Bond subscale measures the extent to which a client and therapist possess "mutual trust, acceptance, and confidence" (p. 224).

Each WAI subscale is scored on a 7-point Likert-type scale ranging from 1 (*never*) to 7 (*always*) and has 12 nonoverlapping items. Subscale scores can range from 12 to 84 and can, if desired, be summed to obtain a total score. Thus, total scores can range from 36 to 252. Higher scores reflect more positive ratings of WA.

Three versions of the WAI are available: a client version, a therapist version, and an observer version. Internal consistency estimates of the three subscale scores, based on initial validation samples of 29 and 25 actual clients and their therapists, ranged from .85 to .92 (client version) and .68 to .87 (therapist version; Horvath & Greenberg, 1986, 1989; for the actual unpublished studies, see Horvath, 1981; Moseley, 1983). Internal consistency estimates of the total scores were .93 (client version) and .87 (therapist version; Horvath & Greenberg, 1986, 1989). Interrater reliability estimates of the observer version were not reported in the original publication. However, internal consistency and interrater reliability estimates of the total scores of this version were, based on initial validation ratings by six doctoral students in counseling or clinical psychology, .98 and .92, respectively (Tichenor & Hill, 1989).

WAI-S

The WAI-S is a 12-item self-report measure of WA (Tracey & Kokotovic, 1989). It, too, has three subscales: Goals, Tasks, and Bond, each of which measure constructs identical to the like-named WAI subscales described above. Items that loaded highest on each of the subscales were retained from the WAI to form the WAI-S.

Each WAI-S subscale is scored on a 7-point Likert-type scale ranging from 1 (*never*) to 7 (*always*) and has four nonoverlapping items. Subscale scores can range from 4 to 28 and can, if desired, be summed to obtain a total score. Thus, total scores can range from 12 to 84. Higher scores reflect more positive ratings of WA.

Two versions of the WAI-S are available: a client version and a therapist version. Internal consistency estimates of the three subscale scores, based on an initial validation sample of 124 pairs of actual clients and their therapists, ranged from .90 to .92 (client version) and .83 to .91 (therapist version; Tracey & Kokotovic, 1989). Internal consistency estimates of the total scores were .98 (client version) and .95 (therapist version; Tracey & Kokotovic, 1989).

Given the prominence WA has achieved in the psychotherapy literature and the widespread use of the WAI and WAI-S, the reliability of their scale scores deserves careful consideration and close empirical scrutiny (Hill & Williams, 2001; Wampold, 2001). Thus, an RG study is, we believe, justified and sorely needed at this time. Our study attempted to meet this need. Specifically, we attempted to answer two basic questions: (a) To what extent are WAI and WAI-S scale scores reliable, generally speaking? and (b) What factors, if any, affect the reliability of these scores?

Method

Sample of Studies, Reliability Estimates, and Study Characteristics

Twenty-five studies were included in this RG study. The studies, published between 1989 and 2002, were located in 11 different journals. The inclusion criteria, literature search procedures, and data-coding procedures are described below.

Inclusion criteria. For a study to be included, three broad guidelines, or inclusion criteria, had to be met. They were the following: (a) The instrument, in this case the WAI or WAI-S, had to be used in the study; (b) the study had to be published in a journal; and (c) the study had to report at least one reliability estimate for at least one of the scale scores for the data in hand.

Literature search procedures. The PsycINFO and ERIC databases were searched, using the term *Working Alliance Inventory*, the acronym *WAI*, and the acronym *WAI-S*. The search was delimited to include only English-language documents. This literature search procedure resulted in 391 matches. Of these 391 matches, 140 (36%) met the first inclusion criterion, that is, used the WAI or the WAI-S. The other 251 were excluded, of which 5 used the Supervisory WAI and one the Advisory WAI. Also, 245 were entirely unrelated to this study (e.g., they used the Wife Abuse Inventory, they were related to the Web Access Initiative, or WAI was the name of one of the authors).

Of the remaining 140 matches, 60 (43%) met both the first and the second inclusion criteria, that is, used the WAI or the WAI-S and were published in a journal. The other 80 were excluded, of which there were 75 dissertations, 3 book chapters, and 2 edited books.

Of the remaining 60 matches, 23 (38%) met all three inclusion criteria, that is, used the WAI or the WAI-S, was published in a journal, and reported at least one reliability estimate for at least one of the scale scores for the data in hand. The other 37 were excluded, all of which failed to report at least one reliability estimate for the study sample.

As an added step, the references of Horvath and Symonds (1991) and Martin et al. (2000), two meta-analyses that included the WAI, were cross-checked with the database matches. Any reference that was not included in the database search was located and read to determine its eligibility. This step resulted in the identification of two more studies that met all three inclusion criteria, increasing the overall total to 25. These 25 studies were, from a percentage standpoint, comparable to, if not slightly higher than, those reported in other published RG studies (Thompson, 1994; Thompson & Vacha-Haase, 2000; Vacha-Haase et al., 1999). These studies were subsequently coded.

Data-coding procedures. To standardize the coding procedure, a data-coding sheet was developed for use by three raters. (The coding sheet is available from the first author.) This allowed the raters to code systematically the reported reliability estimates and study characteristics for each of the 25 studies. Initially, the first two authors coded the studies independently. They were then coded by a third rater, a 4th-year graduate student who was otherwise unconnected with the study. The percentage agreement among the three raters was, out of all possible ratings, 99%. The eight identified discrepancies were discussed, consensus was reached, and minor corrections were made.

For each study, reliability estimates, including internal consistency estimates and interrater reliability estimates, were coded. Two of the 25 studies reported only the range of WAI scale score reliabilities. For these 2 studies, reliability estimates were coded, for the subscale and total scores, as the average of the upper and lower values of the range.

Twelve study characteristics were also coded, including sample size, for both client and therapist; client age; client and therapist gender homogeneity; client and therapist ethnicity; therapist educational level and years of experience; type of client (e.g., volunteer vs. actual client); type of treatment; and session number/time of WAI administration. These study characteristics are typical of those included in previous RG research (see Henson, Kogan, & Vacha-Haase, 2001) and are linked theoretically, if not empirically, to WA (Horvath & Luborsky, 1993; Horvath & Symonds, 1991; Martin et al., 2000).

These data-coding procedures resulted in a total of 73 reliability estimates, including 67 internal consistency estimates, 6 interrater reliability estimates, and (because 8 of the 12 study characteristics were not reported consistently in the majority of studies) 4 study characteristics being analyzed. The 4 study characteristics included in the analyses were sample size for both client and therapist, and client and therapist gender homogeneity, coded as percentage of majority for each.

The typical study in our sample had the following characteristics:

- 56 actual clients (*SD* = 35)—73% female and 27% male of unknown age, 83% European American and 17% of unknown ethnicity, with unknown presenting problem(s);
- 26 therapists (*SD* = 31)—70% female and 30% male of unknown age, ethnicity, and experience level;
- variety of treatment approaches and settings; and

- WAI or WAI-S administered after an unknown or an indeterminable number of sessions.

Data analysis involved characterizing the 73 reliability estimates and the factors, or study characteristics, that affected their variability. Due to the relatively small number of reported reliability estimates, which ranged from 2 to 13 per cell, only descriptive statistics and simple bivariate correlation analyses were conducted. More complex analyses, such as multiple regression, were not feasible.

Results

Descriptive statistics of internal consistency estimates and interrater reliability estimates of WAI and WAI-S scale scores are listed in Table 1. For the client version of the WAI, internal consistency estimates ranged from .77 to .92 (*M* = .87, *SD* = .05, *n* = 9) for the Goals scale scores, .82 to .92 (*M* = .87, *SD* = .03, *n* = 8) for the Tasks scale scores, and .84 to .92 (*M* = .89, *SD* = .03, *n* = 8) for the Bond scale scores. The estimates ranged from .83 to .97 (*M* = .93, *SD* = .04, *n* = 13) for the total scores of this version.

For the therapist version, internal consistency estimates ranged from .87 to .93 (*M* = .90, *SD* = .02, *n* = 6) for the Goals scale scores, .82 to .92 (*M* = .87, *SD* = .05, *n* = 5) for the Tasks scale scores, and .68 to .92 (*M* = .84, *SD* = .10, *n* = 5) for the Bond scale scores. The estimates ranged from .84 to .95 (*M* = .91, *SD* = .05, *n* = 5) for the total scores of this version.

For the observer version, internal consistency estimates ranged from .94 to .98 (*M* = .97, *SD* = .02, *n* = 3) for the total scores. Interrater reliability estimates ranged from .62 to .92 (*M* = .79, *SD* = .12, *n* = 6).

For the client and therapist versions of the WAI-S, mean internal consistency estimates ranged from .92 to .98 (*M* = .95, *SD* = .03, *n* = 3) and .90 to .95 (*M* = .93, *SD* = .04, *n* = 2), respectively, for the total scores.

Figure 1 and Table 2 display box-and-whisker plots and stem-and-leaf plots, respectively, of reliability estimates of WAI and WAI-S scale scores. Visual inspection of Figure 1 shows that reliability estimates of 11 of the 12 scale scores are, regardless of perspective (i.e., client, therapist, or observer version), relatively high (*M*s range from .84 to .97), normally distributed, and minimally variable (*SD*s range from .02 to .10), indicating that they are more or less stable, at least with respect to the studies sampled. The one exception is the WAI total score (observer version, interrater agreement). For this scale score, mean reliability estimates are lower (*M* = .79) and two to three times more variable than they are for the other scores (*SD* = .12). This, however, is not surprising, given that reliability estimates of this version are based on interrater agreement, whereas reliability estimates of the other versions are based on internal consistencies.

Table 1
Descriptive Statistics of Internal Consistency Estimates and Interrater Reliability Estimates of the Working Alliance Inventory (WAI) and Working Alliance Inventory–Short (WAI-S) Scale Scores

Version	*n*	*M*	*SD*	Range
Client				
Goals	9	.87	.05	.77-.92
Tasks	8	.87	.03	.82-.92
Bond	8	.89	.03	.84-.92
Total	13	.93	.04	.83-.97
Client-Short				
Total	3	.95	.03	.92-.98
Therapist				
Goals	6	.90	.02	.87-.93
Tasks	5	.87	.05	.82-.92
Bond	5	.84	.10	.68-.92
Total	5	.91	.05	.84-.95
Therapist-Short				
Total	2	.93	.04	.90-.95
Observer				
Total	3	.97	.02	.94-.98
Total[a]	6	.79	.12	.62-.92

a. Interrater reliability estimates.

Figure 1 also shows that reliability estimates for the total scores are, on average, higher (*M*s range from .91 to .95) for client and therapist versions of the WAI and WAI-S than they are for the Goals, Tasks, and Bond subscale scores (*M*s range from .84 to .90). This is also not surprising given that total scores are based on more items than are subscale scores.

Visual inspection of Table 2 shows that, overall, reliability estimates of WAI and WAI-S scale scores are acceptably high with a modal mean estimate of .92 and, again, are more or less normally distributed.

Intercorrelations among study characteristics, internal consistency estimates, and interrater reliability estimates of WAI and WAI-S scale scores are listed in Table 3. Only two pertinent correlations were statistically significant. The number of clients was correlated negatively with reliability estimates of WAI total scores (observer version, interrater agreement; $r = -.91$, $p < .05$, effect size (es) = .83, $n = 6$), and the number of therapists was correlated negatively with reliability estimates of WAI total scores (observer version, internal consistency; $r = -1.0$, $p < .05$, $es = 1.00$, $n = 3$).

It is difficult to interpret the meaning of these two “significant” findings, however, because the correlations are based on sample sizes of only 6 and 3. Consequently, the results may simply be artifactual. Given this caveat, the first significant correlation suggests that smaller numbers of clients are associated with greater agreement across nonparticipant observers/raters. This

Table 2
*Stem-and-Leaf Plot of Reliability Estimates of the Working Alliance Inventory (WAI) and Working Alliance Inventory–Short (WAI-S) Scale Scores (*n *= 73)*

Stem	Leaf											
.99												
.98	8	8	8									
.97	7											
.96	6	6										
.95	5	5	5	5	5	5						
.94	4	4	4									
.93	3	3	3									
.92	2	2	2	2	2	2	2	2	2	2	2	2
.91	1	1	1	1	1	1	1	1				
.90	0	0	0	0								
.89	9											
.88	8	8	8	8								
.87	7	7	7	7	7	7						
.86												
.85	5	5	5	5	5	5	5					
.84	4	4										
.83	3	3	3									
.82	2	2										
.81	1											
.80												
.79												
.78	8											
.77	7											
.76												
.75												
.74												
.73												
.72												
.71												
.70	0											
.69												
.68	8											
.67												
.66												
.65												
.64												
.63												
.62	2											
.61												
.60												

may be due in part to the difficulties inherent in rating large numbers of people. The other significant correlation cannot, in our judgment, be meaningfully interpreted (*n* = 3).

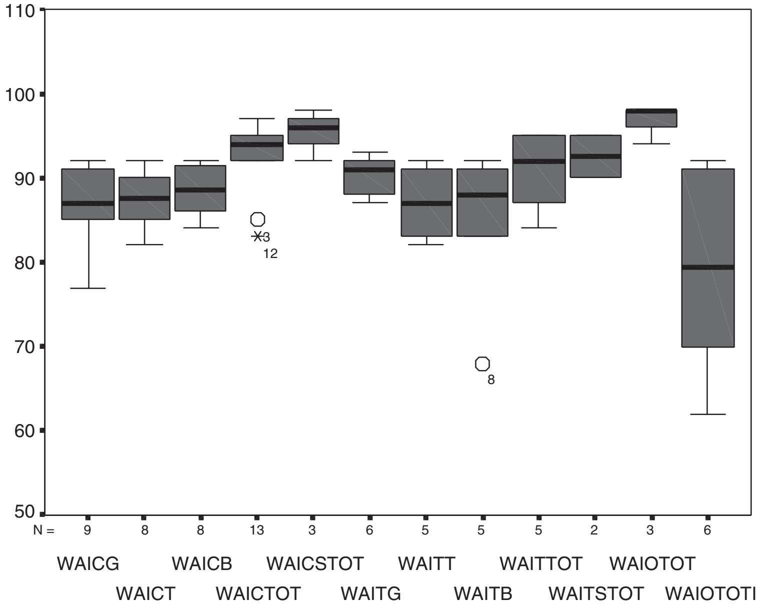

Figure 1. Box-and-whisker plots of reliability estimates of the Working Alliance Inventory (WAI) and Working Alliance Inventory–Short (WAI-S) scale scores.

Note. WAICG = Goals subscale (client version), WAICT = Tasks subscale (client version), WAICB = Bond subscale (client version), WAICTOT = Total (client version), WAICSTOT = Total (client version–short), WAITG = Goals subscale (therapist version), WAITT = Tasks subscale (therapist version), WAITB = Bond subscale (therapist version), WAITTOT = Total (therapist version), WAITSTOT = Total (therapist version–short), WAIOTOT = Total (observer version, internal consistency), and WAIOTOTI = total (observer version, interrater agreement).

Discussion

In the introduction, we raised two basic questions: (a) To what extent are WAI and WAI-S scale scores reliable, generally speaking? and (b) What factors, if any, affect the reliability of these scores? The answer to the first question is relatively clear and straightforward. Based on this particular sample of studies, the answer is "to a great extent," as WAI and WAI-S score reliability estimates were uniformly high, with means ranging from .79 (WAI observer version, interrater reliability) to .97 (WAI observer version, internal consistency). These estimates easily meet professional standards of acceptability (see Cicchetti, 1994). This is especially true of reliability estimates of WAI and WAI-S total scores (*M*s ranged from .91 to .97), which, with one exception (*M* = .79), were even higher than they were for the Goals, Tasks, and Bond subscale scores (*M*s ranged from .84 to .90). Moreover, the reliability estimates varied only minimally across different samples (*SD*s ranged from .02 to .12), suggesting that they are relatively stable. Consistent with these

Table 3
Intercorrelations Among Study Characteristics, Internal Consistency Estimates, and Interrater Reliability Estimates of the Working Alliance Inventory (WAI) and Working Alliance Inventory–Short (WAI-S) Scale Scores

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	.57	–.26	.21	–.52											
n	9	9	9	9											
6	.25	–.28	–.25	–.11	.78*										
n	8	8	8	8	8										
7	–.25	–.08	–.36	.50	.05	.60									
n	8	8	8	8	8	8									
8	–.04	–.28	.28	.16	–.03	.06	.39								
n	13	13	13	13	6	6	6								
9	–.27	–.69	.76	–.47	—	—	—	—							
n	3	3	3	3	1	1	1	0							
10	.57	.61	.25	.28	.06	.17	.10	–.19	—						
n	6	6	6	6	6	5	5	3	1						
11	.71	.47	.00	.00	.26	.35	.22	–.50	—	–.96*					
n	5	5	5	5	5	5	5	3	1	5					
12	.72	.82	.44	.31	.36	.49	.80	–.37	—	.60	.61				
n	5	5	5	5	5	5	5	3	1	5	5				
13	.43	.24	–.65	.14	–.54	–.37	.52	.58	—	.84	.62	.72			
n	5	5	5	5	3	3	3	4	0	3	3	3			
14	.59	–.50	–1.0*	–.25	—	—	—	—	—	—	—	—	1.0*		
n	3	3	3	3	1	1	1	2	0	1	1	1	3		
15	–.91*	.36	.06	.22	—	—	—	—	—	—	—	—	—	—	
n	6	5	6	6	0	0	0	1	0	0	0	0	1	1	

Note. 1 = number of clients, 2 = homogeneity of client gender, 3 = number of therapists, 4 = homogeneity of therapist gender, 5 = Goals subscale (client version), 6 = Tasks subscale (client version), 7 = Bonds subscale (client version), 8 = total (client version), 9 = WAI-S total (client version), 10 = Goals subscale (therapist version), 11 = Tasks subscale (therapist version), 12 = Bond subscale (therapist version), 13 = total (therapist version), 14 = total (observer version), and 15 = total (observer version); 5 through 14 are internal consistency estimates, and 15 is interrater reliability estimates. WAI-S total (therapist version) is not included in the correlation matrix because only two reliability estimates were reported.
*Correlation is statistically significant at the .05 level (two-tailed).

findings, comparable mean reliability and variability estimates were reported by Horvath and Symonds (1991) and Martin et al. (2000).

The answer to the second question, however, is less clear and not as straightforward. Although at least two factors (i.e., the number of clients and therapists) were correlated significantly with mean reliability estimates of the WAI total scores (observer versions: $r = -.91$, $p < .01$, $es = .83$, $n = 6$; $r = -1.0$, $p < .01$, $es = 1.00$, $n = 3$), it simply is not possible, due to notably small sample sizes, to predict which study characteristics are associated with measurement error in WAI or WAI-S scale scores. The correlation analyses reported here unfortunately shed little light on the relationship between WAI and WAI-S reliability estimates and study characteristics. One reason for this is that, as mentioned previously, the majority of studies did not report reliability estimates, thus reducing the sample size (from a possible 60 studies to 25) and the number of study characteristics (from a possible 12 to 4) that could be studied.

Consistent with prior recommendations, it is recommended that future research report reliability estimates more consistently (Vacha-Haase et al., 1999; Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999). Moreover, it is recommended that psychotherapy process researchers report study characteristics in much greater detail (Vacha-Haase et al., 2000). Some study characteristics that could be reported are applicable to both clients and therapists, for example, age, gender, ethnicity, and educational level. Others, however, are more specific to one or the other, for example, socioeconomic status of the client, nature of the client's presenting problem(s), client's expectation for change, client's motivation or readiness for treatment, experience level of the therapist, and therapist's theoretical orientation or allegiance to a particular therapeutic approach. Still others do not necessarily relate to either, for example, the nature of the treatment, integrity of the treatment, treatment setting itself, and specific information about when the instrument was administered.

Given that WA and its measurement may be affected by, among other things, the type of treatment (Cecero, Fenton, Frankforter, Nich, & Carroll, 2001), the time of administration (in terms of number of sessions; Horvath & Luborsky, 1993; Tracey & Kokotovic, 1989), and the perspective measured (i.e., client, therapist, or observer; Fenton, Cecero, Nich, Frankforter, & Carroll, 2001; Tichenor & Hill, 1989), these types of study characteristics, in particular, should be reported more explicitly in future studies. The description of study characteristics provided by Tokar, Hardin, Adams, and Brandel (1996) is exemplary. Their study may be used as a model or template for responsible reporting practices.

As noted by Crocker and Algina (1986), "a test is not reliable or unreliable. Rather, reliability is a property of the scores on a test for a particular group of examinees" (p. 144). Hopefully, researchers will remember this dictum and will begin to write more accurately about test score reliability and, likewise, will begin to report reliability estimates and study characteristics more completely in their scholarly writing. This type of reporting practice will, after all, be beneficial to everyone, especially to those who study the WA and its role in the psychotherapy change process.

References

*Andrusyna, B. A., Tang, T. Z., DeRubeis, R. J., & Luborsky, L. (2001). The factor structure of the Working Alliance Inventory in cognitive-behavioral therapy. *Journal of Psychotherapy Practice and Research*, *10*, 173-178.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research, and Practice*, *16*, 252-260.

*Burkard, A. W., Ponterotto, J. G., Reynolds, A. L., & Alfonso, V. C. (1999). White counselor trainees' racial identity and working alliance perceptions. *Journal of Counseling and Development*, *77*, 324-329.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60*, 236-254.

*Castonguay, L. G., Goldfried, M. R., Wiser, S., Raue, P. J., & Hayes, A. M. (1996). Predicting the effect of cognitive therapy for depression: A study of unique and common factors. *Journal of Consulting and Clinical Psychology*, *64*, 497-504.

*Cecero, J. J., Fenton, L. R., Frankforter, T. L., Nich, C., & Carroll, K. M. (2001). Focus on therapeutic alliance: The psychometric properties of six measures across three treatments. *Psychotherapy*, *38*, 1-11.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*, 481-489.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

*Fenton, L. R., Cecero, J. J., Nich, C., Frankforter, T. L., & Carroll, K. M. (2001). Perspective is everything: The predictive validity of six working alliance instruments. *Journal of Psychotherapy Practice and Research*, *10*, 262-268.

*Florsheim, P., Shotorbani, S., Guest-Warnick, G., Barratt, T., & Hwang, W. (2000). Role of the working alliance in the treatment of delinquent boys in community-based programs. *Journal of Clinical Child Psychology*, *29*, 94-107.

*Forchuk, C. (1995). Development of nurse-client relationships: What helps? *Journal of the American Psychiatric Nurses Association*, *1*, 146-151.

Gaston, L. (1990). The concept of the alliance and its role in psychotherapy: Theoretical and empirical considerations. *Psychotherapy*, *27*, 143-153.

Greenson, R. R. (1965). The working alliance and the transference neurosis. *Psychoanalyis Quarterly*, *34*, 155-181.

Hartley, D., & Strupp, H. (1983). The therapeutic alliance: Its relationship to outcome in brief psychotherapy. In J. Masling (Ed.), *Empirical studies of psychoanalytic theories* (Vol. 1, pp. 1-27). Hillsdale, NJ: Erlbaum.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement*, *61*, 404-420.

Hill, C. E., & Williams, E. N. (2001). The process of individual therapy. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 670-710). New York: Wiley.

Horvath, A. O. (1981). *An exploratory study of the working alliance: Its measurement and relationship to outcome*. Unpublished doctoral dissertation, University of British Columbia, Vancouver, Canada.

Horvath, A. O., & Greenberg, L. S. (1986). Development of the Working Alliance Inventory. In L. S. Greenberg & W. M. Pinsoff (Eds.), *The psychotherapeutic process: A research handbook* (pp. 529-556). New York: Guilford.

*Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*, 223-233.

Horvath, A. O., & Greenberg, L. S. (Eds.). (1994). *The working alliance: Theory, research, and practice*. New York: Wiley & Sons.

Horvath, A. O., & Luborsky, L. (1993). The role of the therapeutic alliance in psychotherapy. *Journal of Consulting and Clinical Psychology*, *61*, 561-573.

Horvath, A. O., & Symonds, D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, *38*, 139-149.

*Hoyt, W. T. (2002). Bias in participant ratings of psychotherapy process: An initial generalizability study. *Journal of Counseling Psychology*, *49*, 35-46.

*Kanninen, K., Salo, J., & Punamäki, R. (2000). Attachment patterns and working alliance in trauma therapy for victims of political violence. *Psychotherapy Research*, *10*, 435-449.

*Kivlighan, D. M., Jr., Clements, L., Blake, C., Arnzen, A., & Brady, L. (1993). Counselor sex role orientation, flexibility, and working alliance formation. *Journal of Counseling and Development*, *72*, 95-100.

*Kokotovic, A. M., & Tracey, T. J. (1990). Working alliance in the early phase of counseling. *Journal of Counseling Psychology*, *37*, 16-21.

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 143-189). New York: Wiley & Sons.

*Long, J. R. (2001). Goal agreement and early therapeutic change. *Psychotherapy*, *38*, 219-232.

Luborsky, L., Crits-Christoph, P., Alexander, L., Margolis, M., & Cohen, M. (1983). Two helping alliance methods of predicting outcomes of psychotherapy. *Journal of Nervous and Mental Disease*, *171*, 480-491.

Marmar, C. R., Weiss, D. S., & Gaston, L. (1989). Towards validation of the California Therapeutic Alliance Rating System. *Journal of Consulting and Clinical Psychology*, *1*, 46-52.

Martin, D. J., Garske, J. P., & Davis, K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *68*, 438-450.

Marziali, E. (1984). Prediction of outcome of brief psychotherapy from therapist interpretive interventions. *Archives of General Psychiatry*, *41*, 301-305.

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology*, *37*, 113-115.

Moseley, D. (1983). *The therapeutic relationship and its association with outcome*. Unpublished master's thesis, University of British Columbia, Vancouver, Canada.

*Multon, K. D., Kivlighan, D. M., Jr., & Gold, P. B. (1996). Changes in counselor adherence over the course of training. *Journal of Counseling Psychology*, *43*, 356-363.

*Patton, M. J., & Kivlighan, D. M., Jr. (1997). Relevance of the supervisory alliance to the counseling alliance and to treatment adherence in counselor training. *Journal of Counseling Psychology*, *44*, 108-115.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

*Raue, P. J., Castonguay, L. G., & Goldfried, M. R. (1993). The working alliance: A comparison of two therapies. *Psychotherapy Research*, *3*, 197-207.

*Raue, P. J., Goldfried, M. R., & Barkhan, M. (1997). The therapeutic alliance in psychodynamic-interpersonal and cognitive-behavioral therapy. *Journal of Consulting and Clinical Psychology*, *65*, 582-587.

*Raue, P. J., Putterman, J. T., Goldfried, M. R., & Wolitzky, D. L. (1995). Effect of rater orientation on the evaluation of therapeutic alliance. *Psychotherapy Research*, *5*, 337-342.

*Rosen, M. I., Desai, R., Bailey, M., Davidson, L., & Rosenheck, R. (2001). Consumer experience with payeeship provided by a community mental health center. *Psychiatric Rehabilitation Journal*, *25*, 190-195.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal*, *13*, 51-59.

Saunders, S. M., Howard, K. I., & Orlinsky, D. E. (1989). The Therapeutic Bond Scales: Psychometric characteristics and relationship to treatment effectiveness. *Psychological Assessment*, *1*, 323-330.

*Sexton, H. (1996). Process, life events, and symptomatic change in brief eclectic psychotherapy. *Journal of Consulting and Clinical Psychology*, *64*, 1358-1365.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

*Solomon, P., Draine, J., & Delaney, M. A. (1995). The working alliance and consumer case management. *Journal of Mental Health Administration*, *22*, 126-134.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195.

*Tichenor, V., & Hill, C. E. (1989). A comparison of six measures of working alliance. *Psychotherapy*, *26*, 195-199.

*Tokar, D. M., Hardin, S. I., Adams, E. M., & Brandel, I. W. (1996). Clients' expectations about counseling and perceptions of the working alliance. *Journal of College Student Psychotherapy*, *11*, 9-26.

*Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the Working Alliance Inventory. *Psychological Assessment*, *1*, 207-210.

*Tryon, G. S., & Kane, A. S. (1995). Client involvement, working alliance, and type of therapy termination. *Psychotherapy Research*, *5*, 189-198.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20.

Vacha-Haase, T., Kogan, L. R., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement*, *61*, 45-59.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*, 509-522.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education*, *67*, 335-341.

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings*. Mahwah, NJ: Erlbaum.

Wilkinson, L., & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Wolfe, B. E., & Goldfried, M. R. (1988). Research on psychotherapy integration: Recommendations and conclusions from an NIMH workshop. *Journal of Consulting and Clinical Psychology*, *56*, 448-451.

Yin, P., & Fan, X. (2000). Assessing the reliability of the Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*, 201-223.

References marked with an asterisk (*n* = 25) indicate studies that were coded and analyzed.