Running Head: COMPARABILITY REVIEW

DRAFT revised January 30, 2004

Comparability of paper and computerized non-cognitive measures: A review and integration

Alan D. Mead

American Institute of Certified Public Accountants

David L. Blitz

Chicago School of Professional Psychology and Illinois Institute of Technology

Paper presented at the eighteenth annual conference of the Society for Industrial and Organizational Psychology, April 2003, Orlando, Florida. Correspondence regarding this paper may be directed to either author: Alan D. Mead, American Institute of Certified Public Accountants, 1230 Parkway Ave., Suite 308, Ewing, NJ 08628, email: amead@aicpa.org. David L. Blitz, Chicago School of Professional Psychology and Illinois Institute of Technology, 47 West Polk Street, Second Floor, Chicago IL 60605, email: blitdav@iit.edu.

ABSTRACT

This study reviewed the literature on equivalence of paper- and computer-administered non-cognitive assessments. We summarize the designs available for analysis of such studies and previous research. Although dozens of studies have been completed, only 6 studies with 41 effects and a total sample size of N=760 met our criteria. Our meta-analysis of these results did not support hypothesized computerization effects.

Comparability of paper and computerized non-cognitive measures: A review and integration

Increasingly, many HR services involving surveys, job-attitude measures and employment testing are being implemented using computers and computer networks such as the Internet. These services are being migrated to automated delivery channels for the same efficiency reasons that drive banks to offer on-line banking and organizations to offer on-line HR benefits information: properly implemented automated services are more available, more convenient, more efficient, and less costly to operate. As more web services are offered, organizations gain even greater economies of scale and cost savings become more assured. Thus this automation process provides greater value. And as a result, more surveys, questionnaires, and tests are being administered by computer and this trend is likely to continue.

But what if the automated measures are not comparable with their paper-and-pencil counterparts? This concern surfaced with the very first computerized assessments (e.g., see Vinsonhaler, Molineaux & Rodgers, 1968) and is ensconced in the American Psychological Association's *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986). The Guidelines state that computerized and paper-and-pencil measures are to be considered comparable only after appropriate research has established this empirically. Specifically:

> When interpreting scores from computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each others, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made

approximately the same by rescaling the scores from the computer mode. (APA, 1986)

This paper has two goals. First, it will review the research designs used to assess comparability and describe the strengths and weaknesses for answering the research question, "Are the paper- and computer-based versions comparable?" Researchers should choose the best design possible. Different designs also have different implications for meta-analysis.

A second goal is to provide an empirical meta-analysis of the comparability of measures other than cognitive ability measures. Cognitive ability measures have previously been reviewed (see Mead & Drasgow, 1993) and relatively frequently studied. Yet non-cognitive measures are probably more often used and encompass a greater diversity (of traits, factor structures, scoring methods, item types, etc.). Thus it is important to review the literature on these non-cognitive measures to determine the accumulated knowledge.

Research on Comparability

Numerous reasons have been proposed to doubt comparability. In the case of computerized surveys, attitude measures, and questionnaires, respondents may have different privacy perceptions regarding computerized and paper forms. Individuals can generally see whether their identity is indicated on paper forms while it is generally unclear whether on-line responses could be traced back to identify themselves; often on-line surveys describe privacy assurances but these may not be trusted by the respondent. Perceptions of anonymity may, for example, be reinforced if the survey does not require an individual login, is available from anywhere, and is hosted by a third party. On the other hand, if the respondent must supply their employee number to access the survey, which is hosted on the company intranet, then even assurances of privacy may not dispel suspicions of skeptical respondents.

Differences between paper- and computer-based tests in responding, such as ability to omit or change answers, has been suggested as a factor affecting comparability (Spray, Ackerman, Reckase, & Carlton, 1989; King & Miles, 1995).

Oswald (19xx) has shown that undergraduate volunteers respond differently … However, Mead and Counssins-Read (19xx) used a different design and found very different results…

Also, respondents using computers may have a different experience due to lack contact with an administrator. For example, it is common for respondents to personality questionnaires to have occasional questions. In paper testing, these individuals may receive clarification of the instructions from an administrator. These individuals would have to proceed on an automated test with their idiosyncratic understanding of the instructions.

Further, given that web-based measures may often be accessible from anywhere, respondents to computerized measures may complete the measures in an inappropriate (or, at least, non-standardized) time or place. Mead and Coussons-Read (2002) reported that "a surprising number" of respondents to a web-based personality measure completed the instrument during the late night and early morning hours.

The aggregate effects of such issues seem sufficient to exercise caution regarding the comparability of paper and computerized forms. And comparability is important for I/O psychologists in at least two situations. First, many testing programs require paper forms for some uses. For example, paper booklets may be used during occasional computer failures, when group testing precludes the use of computers, or when testing in environments without computers (e.g., job fairs, factory floors, etc.). It is also very common for the validity evidence and other documentation of an instruments psychometric properties to be based on a paper version of the instrument.

Previous reviews

[This section is currently under construction]

Comparability Research Designs

Several research designs have been chosen by researchers in this area. In this section, we describe some of the strengths and limitations of designs. We will also examine the degree to which different designs can be meta-analyzed and the data required in original reports to allow inclusion in a meta-analysis.

Table 1 presents a simple taxonomy of study designs. The broadest difference between designs is whether they are between-subjects or within-subjects.

Between-Subjects Designs. Between-subjects designs are those in which different groups of participants respond to the paper and computerized measures. Typically, the means of the paper and computerized measures are compared (using t-tests, ANOVA/ANCOVA, etc.) and significant differences are taken as an indication of non-equivalence. Such studies seem to follow directly from typical experimental designs such as drug studies and are readily "meta-analyzable." An obvious methodological issue is the equivalence of the groups; if the groups are not formed on the basis of random assignment then any results could be caused by the medium of administration, the group differences, or some interaction of these two factors. This is particularly a concern if the different samples of convenience define the computer- and paper-administration groups. For example, if the tests were administered on computer to workers at a new facility and on paper to workers at an older facility, any effect found could well be due to differences in the characteristics, context, administration procedures, etc. at these two facilities rather than the computerization.

Even if the two groups are randomly equivalent, between-subjects designs do not directly test the hypothesis of measurement equivalence because it is possible for the *mean* test scores of the computerized and paper versions to be exactly equal while the *rank-order* of the individuals in the groups entirely different. In such a situation, the versions are definitely not equivalent but because the between-subjects design does not evaluate the equivalence of the rank-orderings of the versions, no difference would be found. Similarly, it would be possible for the ran-order of individuals in the two groups to be identical while the mean scores could be very different. Because between-subjects designs cannot assess similarity of rank-orderings in the two groups, they are a very incomplete test of equivalence.

Another significant methodological issue is more subtle; mean comparisons such as ANOVA are not designed to test equivalence. Standard statistical techniques developed to compare experimental and control groups always test non-equivalence. And in the logic of standard hypothesis testing, a failure to reject the null hypothesis of no differences is not at all the same as demonstrating the null hypothesis. Thus equivalence of the paper and computerized groups is typically "shown" by failing to reject the null hypothesis of no differences but this is a misuse of the statistical methods (see Cohen, 1994).

A primary result of such misuse is that the probabilities of Type I and II errors changes radically when the statistical tests are used this way. Studies that misuse significant testing in this way *must* have a high degree of power; otherwise, the study is likely to find equivalence regardless of the true equivalence or non-equivalence. The primary influence on power is these designs is sample size and thus such studies must have large sample sizes in order to produce results that are at all meaningful.

As an alternative to traditional statistical hypothesis testing in between-subjects designs, Rogers, Howard and Vessey (1993) present a framework in which familiar hypothesis testing

methods can be used to test equivalence hypotheses. The method requires researchers to make a judgment of the degree of difference between the computerized and paper groups that would constitute non-equivalence. A pair of one-sided significant tests are then performed in such a way that if both one-sided null hypotheses are rejected then equivalence can be concluded. An obvious issue with this procedure is choosing the degree of difference that should be considered important—different choices (perhaps for different purposes) may well result in *opposite* conclusions. Such a choice probably defies a simple 5% rule, as is commonly used to set alpha is most significance tests. This method can also be incorporated into a meta-analytic framework (see Rogers, et al., 1993 for an example).

Two other statistical frameworks exist as alternatives (Reise, Widaman, & Pugh, 1993): structural equations modeling (SEM) and item response theory (IRT)[1]. The SEM approach assumes that several scales are investigated and tests the equivalence of the intercorrelation matrix of the scales for the paper and computerized conditions. This approach has several advantages over traditional approaches. It seems unlikely that medium of administration could lead to radical changes in rank ordering of individuals and still not effect the correlation of that scale with other scales. Thus the SEM approach may well have greater sensitivity to changes in rank-order due to the administration medium. Also of interest to any practitioner, it provides a straight-forward means of including criterion-related validity information.

The SEM approach does assume that multiple scales are administered and the number of scales used, the reliability of the studied scales, degree of intercorrelation, and the sample size of the respondents probably influences the results. Fewer, highly-reliable scales, evaluated in large samples are more likely to evidence equivalence (Cheung & Rensvold, 2001). In particular, the sample size requirements for stable modeling is probably significantly higher than is the case with simpler experimental designs. The approach is also more complex to implement and thus

studies employing SEM are harder to review. Finally, the SEM approach may be more sensitive to artifacts due to non-random assignment because it is more likely that the large-scale samples are samples of convenience rather than experimental groups recruited for the study and randomly assigned to paper of computerized treatment groups.

The IRT approach to equivalence takes a different approach. The individual items of the paper and computerized scales are modeled and one of the well-researched differential item functioning (DIF) frameworks (see Raju, van der Linden & Fleer, 1995; Thissen, Steinberg, & Wainer, 1988; Lord, 1980) are used to compare the models for identical items administered on paper or computer. The IRT approach provides a scrutiny of individual items that is not present in the previous methods and thus IRT could be used to address hypotheses about individual items. For example, in studying a test of mixed verbal content, it could be hypothesized that the greater the number of words in an item, the more likely that the medium of administration will result in non-comparability. IRT DIF methods may be relatively robust to differences in the mean ability or attitude or trait between the computer and paper groups (Hambleton & Swaminathan, 1985); in fact, unlike other between-subjects methods, mean differences are explicitly removed by the DIF analysis.

The IRT method does not directly assess the rank-ordering of candidates in the paper and computerized groups. However, if no items are found to exhibit DIF it is highly unlikely that the rank-orders would be disturbed by computerization. This is simply because it is unlikely that all the items of the paper scale could have the same functional relationship with the underlying trait as their computerized counterpart and yet those traits be meaningfully different.

The IRT approach does have some disadvantages. Fitting IRT models to data for DIF analysis requires fairly large samples (see Chauh, Drasgow, & Leucht, 2002) and most common IRT models assume that a single factor underlies the item responses (i.e., that each test measures

a single "thing"). In addition to many people, fitting IRT models may require many items. Lord (1968) suggested having at least 50 items and the present authors have had mixed result on short scales of 10-15 items. Many measures used by practitioners are probably not amenable to analysis using IRT because they are too multi-dimensional and short. In cases where IRT is not appropriate, it may be possible to use SEM to examine individual items (see Raju, Laffitte & Byrne, 2002). Also, different DIF methods can give quite different results (see Chuah, Drasgow & Roberts, 2002).

Within-Subjects Designs. Within-subjects designs are those in which the same individuals take both the paper and computerized measure (or possibly alternate forms of the measure). Typically, the scores on the two versions are correlated, allowing a direct examination of the hypothesis that participants will be ranked in the same order by the two versions. Of course, the arguments listed above against using standard hypothesis testing apply. One approach is to report effect sizes and confidence intervals.

It is important to counter-balance the order of administration to ensure that any order effects do not affect the results. Also, most of the issues affecting test-retest reliability studies apply: If the same test form is administered twice without appreciable delay then the results will be contaminated to an unknown degree by a practice effect (for ability forms) or by participant's desire to answer consistently on the two occasions or their contemplation of personality of attitude items (for non-ability forms; see Hamilton & Shuminsky, 1990; Knowles & Beyers, 1996). If a length of time elapses between administrations, then maturation may affect the results. If alternate forms are used, then any departure from strict parallelism will artificially depress the cross-mode correlation. Split-half-like procedures could be used (where half the measure is administered on computer and half on paper) but there are different ways to split halves and these might result in different outcomes.

In addition to the above considerations, within-subjects designs are less commonly used than between-subjects designs because it is more difficult and time-consuming to arrange for participants to complete both a paper and computerized measure.

Conclusions about designs. Different designs have different strengths and weaknesses. The within-subjects design is necessary in order to directly assess the similarity of the rank-order of test-takers in the two conditions. There are three dominant analysis models for between-subjects designs, "Experimental", SEM and IRT. Although there has been some investigation of combining meta-analysis and SEM (Schmidt, 1992; Viswesvaran & Ones, 1995) there is no way to cumulate SEM results across studies unless the studies all contain identically-structured covariance matricies. Similarly, IRT analyses cannot be meaningfully meta-analyzed.

Method

Literature Search. A literature search was conducted to identify published and unpublished studies comparing paper-and-pencil administrations of noncognitive tests. Several methods were used to obtain validity coefficients and descriptive information for the present study. First, we conducted a computer-based literature search in PsychINFO (1980-2003) and ERIC (1980-2003) using the following filters: "PAPER and COMPUTERIZED". Second, we conducted a manual search in *Computers in Human Behavior* from 1993-2003. Third, we hand-searched conference programs from previous annual conferences of the Society for Industrial and Organizational Psychology (SIOP) and the American Educational Research Association (AERA) for potential articles to be included in the present review. Fourth, we sent requests to researchers whom we knew to publish in this area. Fifth, we conducted Internet searches on Google.com for references to prominent review papers. And finally, we examined the references of obtained papers for additional research. A total of 104 studies were identified in this way. However, most

of these concerned cognitive ability tests or failed to provide needed data. Using the selection criteria outlined below, 6 studies were included.

Criteria for Inclusion. Because there were few applicable studies, we included all within-subjects studies that reported cross-mode correlations and reliabilities for non-cognitive ability measures. We accepted coefficient alpha and test-retest estimates of reliability and we accepted one study that only reported paper-and-pencil reliability (we assumed that the different versions had equal reliability). Six studies containing 41 results and a total sample size of 760 met our criteria. Three results were omitted from these papers—two because they were cognitive ability and a random responding scale with an inappropriate reliability.

Results

Table 3 presents the statistics extracted from the primary studies. The primary focus of this analysis was the disattenuated cross-mode correlation. Disattenuation refers to a statistical correction for unreliability. If computerization had no effect, the expected correlation of paper- and computer-based versions of a test would the square-root of the product of the two versions' reliabilities. Thus disattenuation allows for a simple comparison: Disattenuated values near 1.0 represent no effect of computerization while values below 1.0 index the degree of non-equivalence caused by computerization[2]. The column on the far right of Table 3 is the disattenuated cross-mode correlation which was calculated in the standard way:

$$r^{*}_{cp} = \frac{r_{cp}}{\sqrt{r_{pp} r_{cc}}} \qquad (1)$$

where r^{*}_{cp} is the disattenuated cross-mode correlation, r_{cp} is the observed correlation of the paper and web versions, r_{pp} is the reliability of the paper-based version and r_{cc} is the reliability of the computerized version.

Finally, we computed the mean disattenuated cross-mode correlation. We computed this as a simple average and weighted by sample size, obtaining very close agreement between these two methods (r=1.03 and r=1.02, respectively).

We did not conduct moderator analyses for several reasons. First, the effects are all very close to 1.0. Second, the number of effects is small. And finally, these effects arose from only six studies, casting doubt on the typical assumption of uncorrelated error terms.

Discussion

[This section is currently under construction]

[introductory remarks; lack of moderator analyses; main results; conclusions; limitations and directions for future study]

References

References marked with an asterisk indicate studies included in the meta-analysis.

American Psychological Association. (1986). *Guidelines for Computer-Based Tests and Interpretations*. Washington, DC: Author.

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*, 236-264.

Chuah, S. C., Drasgow, F., & Luecht, R. M. (2002). *How Big is Big Enough? Sample Size Requirements for CAST Item Parameter Estimation*. Manuscript submitted for review.

Church, A. H. (2001). Is there a method to our madness? The impact of data collection methodology on organizational survey results. *Personnel Psychology, 5*, 937-969.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997-1003.

Hambleton, R. H. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hamilton, J. C. & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology, 59*, 1301-1307.

Kim, Jong-Pil. (1999). *Meta-Analysis of Equivalence of Computerized and P&P Tests on Ability Measures.* Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 13-16, 1999).

King, W. C. & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*, 643-651.

Knowles, E. S. & Beyer, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology, 70* (5), 1080-1090.

*Kobak, K. A., Reynolds, W. M., & Greist, J. H. (1993). Development and Validation of a Computer-Administered Version of the Hamilton Anxiety Scale. *Psychological Assessment, 5*, 487-492.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of automated and conventional educational and psychological tests: A review of the literature* (College Board Report No. 88-8). Princeton, NJ: Educational Testing Service.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

*Mead, A. D. & Coussons-Read, M. (2002, April). *The equivalence of paper- and web-based version of the 16PF Questionnaire.* Paper presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

*Peterson, L., Johannsson, V., & Carlsson, S.G. (1996). Computerized testing in a hospital setting: Psychometric and psychological effects. *Computers in Human Behavior, 12*, 339-350.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. F. (2002, April). *Web-based vs. paper and pencil testing: A comparison of factor structures across applicants and incumbents.* Paper presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, CA.

*Potosky, D., & Bobko, P. (1997). Computer versus paper-and-pencil administration mode and response distortion in non-cognitive selection tests. *Journal of Applied Psychology, 82*, 293-299.

Raju, N., S., Laffitte, L., J., & Byrne, B., M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529

Raju, N. S., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*, 353-368.

Richman, W.L., Kiesler, S., Weisband, S., Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754-775.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566.

Rogers, J. L., Howard K. I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553-565.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*(10), 1173-1181.

Stanton, J. M. (1998). An empirical evaluation of data collection using the Internet. *Personnel Psychology, 51*, 709-725.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Tonidandel, S., Quinones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87*(2), 320-332.

Vinsonhaler, J. F., Molineaux, J. E., & Rodgers, B. G. (1968). An experimental study of computer-aided testing. In H. H. Harman, C. E. Helm, & D. E. Loye (Eds.), *Computer-Assisted Testing Conference Proceedings*, November 1966, Princeton, NJ: Educational Testing Service.

*Vispoel, W. P, Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: A comparison of psychometric features and respondent preferences. *Educational & Psychological Measurement, 61*, 461-474.

*Vispoel, W. P. (2000). Computerized versus paper-and-pencil assessment of self-concept: Score comparability and respondent preferences. *Measurement & Evaluation in Counseling & Development, 33*, 130-143.

Viswesvaran, C. & Ones, D.S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48*, 865-885.

Author Notes

We thank the researchers who contributed research, including: Roy B. Clariana, Frederick L. Oswald, Douglas H. Reynolds, Stephen G. Sireci, Edward W. Wolfe

Table 1. A taxonomy of comparability study designs.

Between-Subjects Designs

- ANOVA, ANCOVA, independent samples t-tests
- Item response theory
- Structural equation modeling

Within-Subjects Designs

- paired t-tests
- correlations

Table 2. A "vote-counting" summary of previous research.

Study Information	Construct	Methodology	Comparability Results
King & Miles, 1995	social desirability	SEM	full equivalence
	Machiavellianism	SEM	full equivalence
	equity sensitivity	SEM	partial equivalence
	self-esteem	SEM	partial equivalence
Stanton, 1998	employee survey		
Donovan, et al., 2000	JDI Supervisor	IRT DIF	full equivalence
	JDI Co-Worker	IRT DIF	partial equivalence
Church, 2001	employee survey	hierarchical regression	method accounted for 2.4% of response variability
	employee survey	hierarchical regression	method accounted for 0.6% of response variability
Ployhart, et al., 2002	personality and biodata	SEM	full equivalence
Sireci, et al., 2002	employee survey	MDS	partial equivalence
Chauh, et al., 2002	Neuroticism	IRT DIF	4 of 13 items DIF
	Extraversion	IRT DIF	1 of 10 items DIF
	Openness	IRT DIF	1 of 10 items DIF
	Agreeableness	IRT DIF	1 of 10 items DIF
	Conscientiousness	IRT DIF	3 of 10 items DIF

Table 3. Cross-mode correlations, reliabilities, and sample sizes.

Study	Construct	r_{cp}	r_{cc}	r_{pp}	n	r^{*}_{cp}
Vispoel, et al., 2001	self-esteem	0.88	0.91	0.92	224	0.963
Vispoel, 2000	ES	0.91	0.92	0.90	212	1.000
	SCH	0.89	0.92	0.92	212	0.967
	OSEX	0.91	0.92	0.92	212	0.989
	HON	0.79	0.78	0.82	212	0.988
	MATH	0.94	0.96	0.96	212	0.979
	SELF	0.88	0.96	0.96	212	0.917
	PAB	0.94	0.96	0.97	212	0.974
	VERB	0.92	0.88	0.88	212	1.045
	SSEX	0.87	0.91	0.92	212	0.951
	PROB	0.83	0.87	0.87	212	0.954
	PAR	0.92	0.94	0.93	212	0.984
	PAP	0.91	0.92	0.94	212	0.979
	SV	0.94	0.95	0.95	212	0.989
Potosky & Bobko, 1997	BIDR-IM	0.94	0.87	0.87	174	1.080
	BIDR-SDE	0.92	0.81	0.81	174	1.136
	Unlikely Virtues	0.78	0.72	0.69	174	1.107
	Execution	0.93	0.84	0.84	174	1.107
	Taking Charge	0.94	0.91	0.90	174	1.039
	Concentration	0.95	0.88	0.89	174	1.073
	Composure	0.94	0.87	0.88	174	1.074
	Sustained Attention	0.91	0.83	0.83	174	1.096
	Attention to Detail	0.95	0.89	0.89	174	1.067
	Decisiveness	0.94	0.91	0.91	174	1.033
Mead & Coussons-Read, 2002	Warmth	0.88	0.83	0.83	64	1.060
	Emotional Stability	0.73	0.75	0.75	64	0.973
	Dominance	0.83	0.77	0.77	64	1.078
	Liveliness	0.88	0.82	0.82	64	1.073
	Rule Consciousness	0.92	0.80	0.80	64	1.150
	Social Boldness	0.93	0.87	0.87	64	1.069
	Sensitivity	0.92	0.82	0.82	64	1.122
	Vigilance	0.74	0.76	0.76	64	0.974
	Abstractedness	0.87	0.84	0.84	64	1.036
	Privateness	0.79	0.77	0.77	64	1.026
	Apprehension	0.82	0.79	0.79	64	1.038
	Openness	0.84	0.83	0.83	64	1.012
	Self-Reliance	0.93	0.86	0.86	64	1.081
	Perfectionism	0.88	0.80	0.80	64	1.100
	Tension	0.83	0.78	0.78	64	1.064

Study	**Construct**	$\mathbf{r_{cp}}$	$\mathbf{r_{cc}}$	$\mathbf{r_{pp}}$	**n**	$\mathbf{r^{*}_{cp}}$
Kobak et al, 1993	anxiety	0.92	0.96	0.96	290	0.958
Peterson et al, 1996	Depression	0.81	0.83	0.83	57	0.972

Note: r_{cp} = reported cross-mode correlation; r_{pp} = paper form reliability; r_{cc} = computer form reliability; n = sample size; r^{*}_{cp} = disattenuated cross-mode correlation

[1] Structural equations modeling and IRT have evolved using distinct terminology and different software because the two approaches have different emphases. However there is a growing awareness that these are really two sides f the same coin and multidimensional IRT models may be indistinguishable from non-linear factor analyses. See McDonald (1999) for a discussion.

[2] Although these values are correlations, disattenuated values above 1.0 can be expected because they are estimates. In fact, if computerization has no effect then the true value being estimated is 1.0 and about half of the estimates would be expected to exceed this amount. On the other hand, if the reliability values under-estimate the true reliability of the versions, then the disattenuated cross-mode correlation would be artificially inflated.