Equivalence of Computerized and Conventional Versions of the Beck Depression Inventory-II (BDI-II)

STEFAN E. SCHULENBERG and BARBARA A. YUTRZENKA
University of South Dakota

This study examined the equivalence of the conventional and computerized versions of the Beck Depression Inventory-II (BDI-II), taking into account that computer aversion may negatively impact computer-administered BDI-II scores by elevating them. Participants were 180 psychology undergraduate students from a medium-sized midwestern university. Participants were divided into four experimental groups. Each group was administered the BDI-II twice in various combinations (conventional only, computerized only, conventional and computerized and vice versa). All participants completed measures of computer aversion and computer experience. Participants who received both versions of the BDI-II were also asked to specify their preference for method of administration. Independent samples *t*-test results indicated that the computerized and paper-and-pencil versions of the BDI-II may be considered equivalent in terms of measurement validity. Implications for future research are discussed.

It has long been thought that computers may benefit psychological assessment (Smith, 1963). According to Brown (1984), advancements over the years in computerized technology have implications for the assessment of psychological problems. Currently, computers are "ubiquitous, affordable machines" (Olson-Buchanan & Drasgow, 1999, p. 1) that are easily available to the clinician, and the application of computers to psychological testing has been described as extensive (Johnson, 1984), permanent (Butcher, 1987), and increasing (Tseng, Macleod, & Wright, 1997). Computers have been integrated into psychological assessment to the extent that many popular and frequently used paper-and-pencil instruments have been adapted into a computerized format. However, it is not a simple matter to adapt a paper-and-pencil measure into a computerized version. One must take into account whether the two forms (conventional and computerized) are equivalent.

When one is referring to equivalency between paper-and-pencil and computerized versions of the same test, one is essentially asking whether the instruments represent alternate forms of the test in question (Harrell, Honaker, Hetu, & Oberwager, 1987). Kubinger, Formann, and Farkas (1991) noted that, no matter how fine the adaptation, unanticipated consequences may complicate an instrument's items. According to the American Psychological Association's (1986) *Guidelines for Computer-Based Tests and Interpretations*, equivalence between paper-and-pencil and computerized tests may be determined if "(a) the rank orders of scores of individuals tested in alternative

Current Psychology: Developmental • Learning • Personality • Social Spring 2001, Vol. 20, No. 3, 216–230.

modes closely approximate each other, 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" (p. 18). An example of a question relating to equivalence would be whether a computerized adaptation of a conventional depression instrument somehow introduces a new construct (e.g., computer aversion) to the assessment process. That is, the equivalence in this example is one of construct validity. The importance of construct validity with regard to equivalence has been noted by others (e.g., King & Miles, 1995; Neuman & Baydoun, 1998; Turban, Sanders, Francis, & Osburn, 1989).

George, Lankford, and Wilson (1992) and Dimock and Cormier (1991) noted that the failure to account for individual differences (e.g., computer anxiety) across modalities may be a significant factor in understanding mean differences. Gardner, Discenza, and Dukes (1993) astutely pointed out that computer anxiety plays a major role with regard to individuals who are resistant to using computers. In their review of the literature relating to the construct of computer anxiety, Meier and Lambert (1991) found a number of terms used to describe this construct, including "phobia" and "anxiety." These authors suggested that the term "aversion" should be used when describing this construct so that it is not confused with severe psychological conditions that warrant clinical attention. Thus, in the present study, computer aversion is the term used to reference the construct of computer anxiety.

Many instruments are available for the assessment of computer aversion. Those interested in a comprehensive review are encouraged to peruse LaLomia and Sidowski (1993). These authors provided an in-depth analysis of every computer aversion scale available at the time of their study. LaLomia and Sidowski noted that, of the instruments examined, only the Computer Aversion Scale (CAVS; Meier, 1985, 1988) appeared to be grounded in theory, namely, social learning theory.

Dimock and Cormier (1991) noted that individuals who lack experience with computers, or who are otherwise not familiar with them, will have greater feelings of anxiety when working with a computer than their computer-familiar counterparts. Further, an increase in anxiety could negatively impact a person's performance when being assessed through a computerized format. Nurius (1990) noted that there is a lack of research concerning differences in response patterns as a consequence of anxiety or limited computer familiarity. Brown (1984) pointed out that individuals experiencing symptoms of depression may encounter difficulties with computers. The idea that computer anxiety may hinder a computerized assessment has been noted by others (Ford, Vitelli, & Stuckless, 1996; George et al., 1992; Tseng et al., 1997).

There are many equivalence studies across a wide variety of areas (e.g., personality assessment, intelligence assessment) available in the literature. One area receiving increasing attention in the literature relates to the equivalence of instruments of negative affect. In their review of the literature, Schulenberg and Yutrzenka (1999) concluded that there is a strong argument for the use of computerized adaptations of paper-and-pencil instruments; however, clinicians must be wary because of the complexity of the equivalence issue, particularly as it relates to instruments of negative affect. There are many paper-and-pencil instruments of negative affect available to the practicing clinician in a computerized format. One example of this technological transition is the Beck Depression Inventory. Over the years, a plethora of research has been generated to support the reliability and validity of the paper-and-pencil BDI with a variety of populations (see Beck, Steer, & Garbin, 1988, for a review). There are much fewer data to support the utility of the computerized versions. This is particularly true of the recently developed computerized version of the second edition of the BDI (BDI-II; Beck, Steer, & Brown, 1996). However, despite the paucity of research available in this area, there is some data on previous versions of the BDI. For example, Peterson, Johannsson, and Carlsson (1996) found that scores for the computerized Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) tended to be higher than their conventional counterparts, although this finding was not statistically significant. Peterson et al. noted that increased computer scores may result from the computerized modality of assessment. Further, any disparities between conventional and computerized assessment instruments is a concern, particularly when norms are adapted from paper-and-pencil to computerized tests in order to facilitate interpretation (at a minimum for measures requesting sensitive information, as is so often the case with instruments of negative affect).

The purpose of this study was to systematically examine the equivalence of the written and computerized versions of the BDI-II, taking into account the consideration that computer aversion may negatively impact the results of the BDI-II by elevating computer scores (see, for example, George et al., 1992). Based on the findings of numerous others, it was hypothesized that the written and computerized versions of the BDI-II would be equivalent. However, if the two forms were not found to be statistically equivalent, then it was hypothesized that scores obtained from the computerized version of the BDI-II would be inflated. This potential disparity would be accounted for (at least in part) through the impact of computer aversion. Other potential confounding variables (e.g., computer experience, preference for method of administration) and their relationship to nonequivalence, were also examined.

METHODS

Participants

One hundred and eighty undergraduate students enrolled in psychology courses at a medium-sized midwestern university participated in the study.

Measures

Measures included: (a) a Demographic Data Form, (b) the conventional (written) and (c) the computerized versions of the BDI-II, (d) the Computer Aversion Scale, (e) the Computer Experience Index, and (f) the Preference for Mode of Administration Questionnaire.

Demographic questionnaire. The Demographic Data Form included questions addressing the participants' gender, age, ethnic/racial background, exposure to computer training, and prior experience in psychological research.

Beck Depression Inventory-II-written version (BDI-II). The BDI has been used extensively over the years in measuring depressive symptoms in both clinical and nonclinical populations (Beck, Steer, Ball, & Ranieri, 1996; Beck et al., 1988). The BDI asks 21 questions concerning various characteristics of depression. Each item is answered by circling a number between 0 and 3, with larger numbers indicating greater severity. Scores for each question are summed into an aggregate score ranging from 0 to 63. The BDI is generally self-administered, and takes approximately 5 to 10 minutes to complete (Beck et al., 1988). The BDI-II (Beck et al., 1996) continues to be scored in the same fashion as previous versions (BDI-I; Beck et al., 1961; BDI-IA; Beck & Steer, 1993). Key changes in the BDI-II lie in the alteration of some of the items and the time frame that an individual is asked to consider when responding to items (see Beck et al., 1996, for a description of changes). Steer, Ball, Ranieri, and Beck (1997) pointed out that the BDI-II retains the same number of items (21), with the same range of scores possible. Further, the amount of time that an individual considers when responding to items was expanded from one week to two, to better coincide with criteria for the depressive disorders as outlined in classification systems such as the *Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition* (DSM-IV; American Psychiatric Association, 1994).

Beck, Steer, Ball, et al. (1996) compared the BDI-IA (a revised version of the BDI) and the BDI–II with a sample of 140 outpatients experiencing a variety of psychiatric difficulties (e.g., mood disorders, anxiety disorders). They reported high internal consistencies for both the BDI-IA (α = .89) and the BDI-II (α = .91). Steer et al. (1997) examined the construct validity of the BDI-II as it relates to the various subscales of the Symptom Checklist-90–Revised (SCL-90–R; Derogatis, 1983). They used an outpatient sample of 210 adults (127 women, 83 men) assessed for psychiatric difficulties such as depression and anxiety. Their results indicated high internal consistency (α = .92) for the BDI-II. Support for the validity of the BDI-II was evidenced by Pearson correlations between the BDI-II and the Depression and Anxiety scales of the SCL-90–R (.89 and .71, respectively).

Beck Depression Inventory-II-computerized version (C-BDI-II). The C-BDI-II is a computerized version of the conventionally administered BDI-II, and is available as part of a computerized application called *OPTAIO Provider's Desktop*. This application combines record keeping with an on-line assessment system and is available from The Psychological Corporation. Although a computerized version of the BDI has garnered empirical support in the literature (e.g., Steer, Rissmiller, Ranieri, & Beck, 1994, calculated a coefficient alpha of .92), to the authors' knowledge there were not any published empirical studies directly concerning the computerized BDI-II at the time this investigation was conducted.

Computer Aversion Scale (CAVS). The Computer Aversion Scale (CAVS; Meier, 1985, 1988) is a 31–item scale based on social learning theory that was designed for use with mental health consumers and professionals (LaLomia & Sidowski, 1993; Meier, 1988). The CAVS is described in detail by Meier (1988). Each item asks for a "true" or "false" response. Meier noted that the CAVS yields a Total score (all items), and three theoretical expectancy subscale scores: (a) Efficacy (10 items referring to feelings of competence in order to act to produce a desired consequence), (b) Outcome (10 items referring to expectancies about the kinds of acts that may lead to a desired consequence), and (c) Reinforcement (11 items referring to whether consequences and goals are congruent). The CAVS Total score is determined by totaling the number of responses indicating computer aversion, with higher scores revealing greater aversion.

Meier (1988), administering the CAVS to 270 undergraduates sampled from a large northeastern university, performed a principal components factor analysis with varimax rotation, finding support for the model of computer aversion. He reported alpha coefficients for the Total score (.89), the Efficacy score (.80), the Outcome score (.81), and the Reinforcement score (.74), with means (and standard deviations) for the four scales to be 12.62 (6.72), 3.35 (2.66), 4.78 (2.98), and 4.47 (2.43), respectively. Meier suggested that the reliability and validity of the CAVS was such that the instrument is a useful screening device for computer aversion. Meier suggested using a cutoff score of 19 (one standard deviation over the mean) as an indication that computer aversion is problematic. The factor analysis in Meier's study also suggested a new 10–item scale, termed Negative Feelings for Computers ($M = 2.75$, $SD = 2.72$).

Meier and Lambert (1991) tested the psychometric properties of three computer aversion scales on 1,234 introductory psychology students. The three computer aversion scales involved in the study were the Attitudes Toward Computer Scale (ATC; Kjerulff & Counte, 1984), the Computer Aversion Scale (CAVS; Meier, 1985, 1988), and the Computer Anxiety Rating Scale (CARS; Heinssen, Glass, & Knight, 1987). The Computer Experience Questionnaire (CEQ; Lambert & Lewis, 1989) was used to determine participants' computer exposure. Meier and Lambert's findings indicated that although the scales seemed comparable in terms of reliability (e.g., internal consistency, test-retest), the CAVS had a small edge (more consistent test-retest reliability). Convergent validity scores between the CAVS and the CARS ranged from between .62 to .67, while the correlations between these measures and the ATC were much lower.

Computer Experience Index (CEI). The CEI is a 13–item questionnaire designed specifically for use in this study, and was modeled after instruments purported to measure similar constructs (e.g., Dimock & Cormier, 1991). Examples of items on the CEI include "How often have you programmed a computer to perform a task?" and "How often do you access the internet?". The CEI utilizes a Likert-like response format, with a range from 1 (e.g., "never") to 5 (e.g., "daily"). Higher scores indicate greater levels of computer experience. Potential scores range from a low of 13 to a high of 65.

Preference for Mode of Administration Questionnaire. The Preference for Mode of Administration Questionnaire (PMAQ) was originally designed by Merten (1994) and published by Merten and Ruch (1996). The PMAQ was originally a 9–item questionnaire designed to determine an individual's preference for method of assessment (conventional vs. computer). Each item is scored as a 1 (preference for conventional assessment), a 2 (no preference), or a 3 (preference for computerized assessment). Scores of 2 (no preference) for each item will yield a result of 18, indicating an individual's overall lack of preference. The instrument is interpreted as to whether the obtained score is significantly disparate from the expected total score of 18. Merten and Ruch recommended excluding item 5 ("I found it more tedious to work: on the questionnaire/at the computer/no difference.") from future studies, because psychometrics of the measure are improved (internal consistency .83, split-half reliability .81). In accordance with this, this study used the 8–item version of the PMAQ. As a consequence, the expected score for this study was 16.

Procedures

Participants were recruited from undergraduate psychology courses. Equanimity in the random assignment of participants to one of four experimental conditions was achieved by assembling the questionnaires into four different packets representing the four experimental conditions in the study. To minimize confounds (e.g., order effects), each participant had an equal chance of being in any of the experimental conditions. A counter-balanced design was used such that each group was exposed to one of four methods of administration: (a) the conventional (written, W) version first, followed by the computer version (C), (b) C-W, (c) W-W, and (d) C-C.

Following administration of the first BDI-II, participants completed the CAVS and the CEI. These were followed by the second administration of a written or computerized BDI-II (depending on the experimental condition in question). Participants who completed both versions of the BDI-II were also administered the PMAQ at the end of the study. The average time of completion for the entire battery of tests for all groups was around 15–25 minutes.

RESULTS

Participant Characteristics

Of the 180 participants, 141 (78.3%) were women and 39 (21.7%) were men. Similar proportions of men and women were found for each experimental group. One hundred and seventy eight participants reported their age (two participants in Group 4 did not offer a response), with a range from 18 to 50 (*M* = 21.95; *SD* = 5.47). The age characteristics for the total sample were distributed similarly in each of the four groups. One hundred and sixty five (92%) of the participants reported their race/ethnicity as Caucasian; seven (4%) were Native American, three (2%) were African American, three (2%) were Asian American, and one (1%) was a Hispanic American. Race/ethnicity proportions for each of the four groups were similar to those of the total sample.

Primary Analyses-Establishing Equivalence of Conventional and Computerized BDI-IIs

BDI-II. Table 1 presents the means, standard deviations, and ranges for the BDI-II, by order of administration. Although the range of scores across the four groups is

TABLE 1
BDI-II Means, Standard Deviations, and Ranges, By Order of Administration

	Group 1 *n* = 45	Group 2 *n* = 45	Group 3 *n* = 45	Group 4 *n* = 45	Total by Row *n* = 90
First Administration					
BDI-II					
Conventional					
M	9.13		8.53		8.83
SD	6.74		6.92		6.80
Range	0-35		0-39		0-39
Computerized					
M		10.11		10.07	10.09
SD		9.33		8.93	9.08
Range		0-43		0-40	0-43
Second Administration					
BDI-II					
Conventional					
M		9.56	8.27		8.91
SD		9.20	7.58		8.41
Range		0-43	0-40		0-43
Computerized					
M	8.87			9.29	9.08
SD	6.16			8.51	7.39
Range	0-31			0-38	0-38

considerable, mean scores for the participants suggest the expression of a minimal degree of symptomatology. Statistical calculations were performed using the SPSS computer software package (1996). The first analysis involved an independent samples *t*-test on the first administration (Time 1) of the BDI-II, by modality. That is, Time 1 results were combined for Groups 1 and 3 (the BDI-II written version), and for Groups 2 and 4 (the BDI-II computer version), and then compared. This answered the fundamental question as to whether the modalities of the BDI-II were equivalent in this sample. For the computerized version, the mean was 10.09 and the standard deviation was 9.08, while for the written version, the mean was 8.83 and the standard deviation was 6.80. Equal variances were assumed. Results of the *t*-test were not significant. That is, t (df) = 1.05 (178), $p > .05$, indicating that the two forms of the BDI-II were equivalent in this sample.

To further analyze the data, bivariate correlations between Times 1 and 2 for all four experimental groups were determined. The *r* scores were converted to standard *z* scores using an *r* to *z* transformation table. This revealed whether there are differences among the differences. Basically, the idea was to convert the bivariate *r*'s into standard scores, which were then used as a foundation for comparison to determine if the test-retest reliability was the same when different modalities were used, as when identical modalities were used. Calculated correlations for BDI-II administrations in Groups 1,

TABLE 2
Latin Square Analyses of the Conventional and Computerized BDI-IIs for Order and Modality

Source	*SS*	*df*	*MS*	*F*	*PRE*
Between					
Sequence	31.25	1	31.25	.25	.00
Seq. W/Subjects	11,123.00	88	126.40		
Total Between	11,154.25	89			
Within					
Admin. Type	.94	1	.94	.80	
Order	7.61	1	7.61	6.50*	.07
Admin. x Order	102.96	88	1.17		
Total	111.50				

Note. *$p < .05$.

2, 3, and 4, were determined to be .98, .98, .97, and .98, respectively, between Times 1 and 2. All correlations were significant at $p < .05$. Using an r to z transformation table (Runyon & Haber, 1991, p. 536), the correlations for each of the four experimental groups were converted to standard scores. Standard scores for each of the four groups were determined to be 2.30 for Group 1, 2.30 for Group 2, 2.09 for Group 3, and 2.30 for Group 4. The standard scores for each of the four groups were nearly identical, suggesting that the test-retest reliability was virtually identical for each of the experimental conditions, regardless of the version of the BDI-II administered.

A Latin square analysis was conducted on Groups 1 and 2. This provided mean comparisons on order and modality. The results are presented in Table 2. The order variable was statistically significant at the $p < .05$ level, suggesting that the BDI-II scores decreased significantly from Time 1 to Time 2. The effect size calculated for this order effect was .53.

CAVS, CEI, and PMAQ. Table 3 presents results for the Computer Aversion Scale, the Computer Experience Index, and the Preference for Mode of Administration Questionnaire. CAVS scores overall ranged from 0 to 26 ($M = 8.19$; $SD = 5.41$). The CAVS subscales, Outcome, Efficacy, Reinforcement, and Negative Feelings for Computers, had means (and standard deviations) of 2.88 (2.20), 1.88 (2.04), 3.43 (2.15), and 1.86 (2.34), respectively. CEI scores for the total sample ranged from 15 to 56, with a mean of 37.67 and a standard deviation of 7.36. PMAQ scores (administered to Groups 1 and 2 only) ranged from 9 to 24, with a mean of 19.39 and a standard deviation of 3.41.

Correlation between primary variables. A correlation matrix was performed to further clarify the relation between the measures used in this study. The correlation matrix is presented in Table 4. The BDI-II was entered into the matrix by order of administration because the results of the statistical analysis revealed the written and computer forms to be statistically equivalent. The correlation between first and second administrations of the BDI-II was .98. Of additional interest were the small to moderate negative correlations between computer experience (CEI) scores and the CAVS total scores (-.67), as well as computer experience scores and the CAVS subscale scores (correlations ranged from -.42 to -.70).

TABLE 3
CAVS, CEI, and PMAQ Means, Standard Deviations, and Ranges, By Groups and Total Sample

	Group 1 *n* = 45	Group 2 *n* = 45	Group 3 *n* = 45	Group 4 *n* = 45	Total *n* = 180
CAVS					
Total Score					
M	6.78	8.22	9.31	8.47	8.19
SD	5.14	5.39	6.05	4.84	5.41
Range	1-22	0-22	2-26	1-23	0-26
Outcome					
M	2.38	2.89	3.31	2.96	2.88
SD	2.08	2.19	2.48	1.99	2.20
Range	0-8	0-8	0-8	0-8	0-8
Efficacy					
M	1.40	1.89	2.47	1.76	1.88
SD	1.88	1.93	2.46	1.72	2.04
Range	0-8	0-8	0-10	0-7	0-10
Reinforcement					
M	3.00	3.44	3.53	3.76	3.43
SD	2.18	2.13	2.13	2.15	2.15
Range	0-10	0-10	0-9	0-10	0-10
Negative Feelings for Computers					
M	1.44	1.80	2.16	2.02	1.86
SD	1.99	2.43	2.70	2.19	2.34
Range	0-8	0-10	0-9	0-8	0-10
CEI					
M	40.16	38.78	36.29	35.47	37.67
SD	6.65	7.41	7.08	7.54	7.36
Range	28-55	24-56	19-49	15-50	15-56
PMAQ					
M	19.93	18.84	NA	NA	19.39
SD	3.21	3.56	NA	NA	3.41
Range	10-24	9-24	NA	NA	9-24

Note. The titles of instruments are as follows: CAVS = Computer Aversion Scale; CEI = Computer Experience Index; PMAQ = Preference for Mode of Administration Questionnaire.

Secondary Analyses

Internal consistency of measures. Coefficient alphas were calculated for each instrument to add to the literature on each of the measures, especially in view of the experimental nature of instruments such as the CEI and the CAVS. Coefficient alphas for BDI-II scores (first administration written and first administration computer), CEI scores, and PMAQ scores were determined to be .88 and .91, .77, and .83, respectively. Coefficient alphas for the CAVS Total, Outcome, Reinforcement, Efficacy, and

TABLE 4
Interrcorrelations Between Scales

	1	2	3	4	5	6	7	8	9
1. BDI–II[a]	—	.98**	–.19*	–.27*	.24**	.21**	.24**	.16*	.14
2. BDI–II[b]	—	—	–.18*	–.25*	.25**	.21**	.25**	.17*	.15*
3. CEI	—	—	—	.12	–.67**	–.70**	–.42**	–.59**	–.52**
4. PMAQ	—	—	—	—	–.31**	–.17	–.37**	–.24*	–.32**
5. CAVS–Total	—	—	—	—	—	.90**	.77**	.87**	.89**
6. CAVS–O	—	—	—	—	—	—	.49**	.78**	.73**
7. CAVS–R	—	—	—	—	—	—	—	.46**	.72**
8. CAVS–E	—	—	—	—	—	—	—	—	.82**
9. CAVS–N	—	—	—	—	—	—	—	—	—

Note. All instruments involve an N of 180 except the PMAQ, which has an N of 90. The titles of instruments and letters are as follows: BDI-II = Beck Depression Inventory-II; CEI = Computer Experience Index; PMAQ = Preference for Mode of Administration Questionnaire; CAVS-Total = Computer Aversion Scale Total score; CAVS-O = Computer Aversion Scale Outcome score; CAVS-R = Computer Aversion Scale Reinforcement score; CAVS-E = Computer Aversion Scale Efficacy score; CAVS-N = Computer Aversion Scale Negative Feelings for Computers score.
[a] BDI-II first administration, regardless of method of administration.
[b] BDI-II second administration, regardless of method of administration.
*$p < .05$. **$p < .01$.

Negative Feelings for Computers scores were .87, .74, .68, .77, and .83, respectively. Overall, the coefficient alphas suggest that each of the instruments appear to demonstrate internal consistency.

Post Hoc t-tests. Beck et al. (1996) found mean BDI-II scores for the women to be statistically greater than the men in their college student sample. They called for additional research to examine this issue. An additional independent samples *t*-test was performed post hoc comparing first administration BDI-II scores for the women (*n* = 141, *M* = 9.45, *SD* = 7.98) and the men (*n* = 39, *M* = 9.49, *SD* = 8.28) to determine if a gender difference was present. Gender differences by first BDI-II administration were not found to be statistically significant, *t* (*df*) = -.023 (178), *p* > .05.

Additional post hoc *t*-tests were performed comparing the first administration BDI-II scores within each gender, by modality. For example, first administration BDI-II scores for the women for computer (*n* = 72, *M* = 10.17, *SD* = 9.34) and written versions (*n* = 69, *M* = 8.71, *SD* = 6.23) were compared. The results were not significant, *t* (*df*) = 1.08 (139), *p* > .05. First administration BDI-II scores for the men for computer (*n* = 18, *M* = 9.78, *SD* = 8.18) and written versions (*n* = 21, *M* = 9.24, *SD* = 8.56) were also compared. The results were also not significant, *t* (*df*) = .20 (37), *p* > .05. Thus, the gender differences found by Beck et al. (1996) were not supported in this study.

Independent samples *t*-tests were also performed comparing the means of the remaining instruments by gender. These results are displayed in Table 5. Women and men differed significantly in terms of computer experience, with men tending to have more experience than women. Women also tended to experience greater levels of computer aversion. However, means were not indicative of computer anxiety per se

TABLE 5
Post Hoc Independent Samples t-Tests for CEI, CAVS, and PMAQ By Gender

	Women (n = 141) *M (SD)*	Men (n = 39) *M (SD)*	t (*df*)
CEI	36.74 (6.73)	41.03 (8.61)	-3.30 (178)*
CAVS-Total	8.80 (5.49)	6.00 (4.50)	2.92 (178)*
CAVS-O	3.12 (2.17)	2.03 (2.12)	2.80 (178)*
CAVS-R	3.65 (2.26)	2.64 (1.42)	2.65 (178)*
CAVS-E	2.03 (2.08)	1.33 (1.78)	1.90 (178)
CAVS-N	2.08 (2.46)	1.05 (1.64)	2.46 (178)*
PMAQ**	19.39 (3.50)	19.39 (3.23)	-.004 (88)

Note. The titles of instruments are as follows: CEI = Computer Experience Index; PMAQ = Preference for Mode of Administration Questionnaire; CAVS-Total = Computer Aversion Scale Total score; CAVS-O = Computer Aversion Scale Outcome score; CAVS-R = Computer Aversion Scale Reinforcement score; CAVS-E = Computer Aversion Scale Efficacy score; CAVS-N = Computer Aversion Scale Negative Feelings for Computers score.
$^{*}p$ <.05.
** Women (n = 67), men (n = 23).

for either gender. Women and men did not differ significantly in their preference for mode of administration. That is, those that were administered the PMAQ tended to prefer computerization, regardless of gender.

DISCUSSION

The purpose of this study was to examine the equivalence of the conventional and computerized versions of the BDI-II, taking into account the consideration that computer aversion may elevate computer scores. It was hypothesized that the written and computerized versions of the BDI-II would be statistically equivalent. The results of the independent samples *t*-test for the BDI-II supported this hypothesis. That is, with this sample, the difference in means between the computerized and paper-and-pencil versions of the BDI-II suggested that the two forms may be considered equivalent. This finding is consistent with the majority of the research literature on equivalence (see Schulenberg & Yutrzenka, 1999, for a review).

It was also hypothesized that if the two forms were not found to be equivalent, then scores obtained from the computerized version of the BDI-II would be inflated, possibly due to the impact of computer aversion, computer experience, and preference for method of administration. The findings of this study precluded the need to test the secondary hypothesis. Indeed, the CAVS scores for the sample indicated that computer aversion was not a problem per se. Average CAVS scores were consistently below the cutoff proposed by Meier (1988). In fact, it is interesting to note that mean CAVS scores were lower than Meier's standardization sample. Participants in this study likely experienced low levels of aversion toward computers because they were relatively experienced in terms of computer usage. Moreover, they reported a preference for the computerized administration of the BDI-II. PMAQ scores (M = 19.39) for the total sample suggested that participants were not experiencing computer aversion. The sample was well above the score suggesting a preference for computerized administration (*M* = 16). Retrospectively, this may have been predictable given the amount of computer experience reported by participants. That is, CEI scores and demographic questions indicated that the sample as a whole had a good deal of computer experience. The CEI is a promising instrument because of its reliance on a Likert-like response format. However, because the CEI is such a new instrument, additional research remains to be done before it should be considered for routine use. In populations where computer aversion may be problematic, such as individuals who possess little or no experience with computers, the issue as to whether computer aversion artificially inflates the scores generated by a computerized version of an instrument of negative affect remains a question. In future investigations it would be helpful to initially assess for the presence of computer aversion, computer experience, and preference for method of administration prior to attempting to determine the potential effect on computerized BDI-II scores.

The Latin square analysis suggested an order effect in Groups 1 and 2. That is, scores declined on the second administration of the BDI-II, regardless of the method of administration. It is possible that when taking two versions of a measure of negative affect (such as the BDI-II) in such a short span of time, people tend to become more comfortable, which may influence scores. This could also be the result of testing effects or artifact. The design of this study cannot effectively address this issue, thus warranting further examination. It should also be noted that although the drop in scores from first to second administration was found to be statistically significant, it is not practically significant. That is, BDI-II scores in this study did not tend to decline to the extent where it would influence clinical decision-making. Though it would be interesting to see if the drop in depression scores would remain constant with test-retest intervals larger than 6–15 minutes, the test-retest interval in this study. Determining test-retest reliabilities for the BDI-II is another area that would benefit from additional research. Test-retest reliability comparisons across conventional and computerized versions of the BDI-II would also be another method of determining equivalence and reliability of these forms. Thus, one direction for future research would be to use the research design of this study with larger test-retest intervals, which would add important information regarding the stability of these forms.

Although the primary purpose of the study was to determine the equivalence between written and computerized versions of the BDI-II and to collect some psychometric data on the measures used, additional analyses were performed. For example, the correlation matrix revealed small to moderate negative relationships between the CEI and the CAVS Total score and each of the CAVS subscales. That is, the more computer experience one has, the less computer aversion that person will tend to experience, and vice versa. Examination of gender differences on the CAVS and CEI scores suggested that the men in this sample possessed more experience with computers, and felt less aversion toward computers, than women. Again, it should be noted that although these latter differences were found to be statistically significant, they did not appear to be practically significant. For example, the mean CAVS scores for both the men and the women in this sample were not computer averse, so in essence, the gender difference simply reflects that while both were not computer averse, men were slightly less so.

Post hoc analyses suggest that contrary to previously reported gender differences in BDI-II scores (Beck et al., 1996), the BDI-II scores of the men and women in this study did not differ. More research is needed to determine whether gender differences in BDI-II scores are a genuine concern to the interpretive process.

Limitations of This Study

Throughout the study, it was readily clear that students commonly recognized they were given either two identical forms of the BDI-II, or two somewhat different forms of the same questionnaire. Thus, all answers to the second administration of the BDI-II need to be interpreted cautiously because students may have remembered their responses from the previous administration. Limitations related to possible memory effects were magnified because students finished the questionnaires much faster than was anticipated during the initial design of the project, allowing approximately 6–15 minutes between BDI-II administrations. In particular, the results of the second administration of the written BDI-II should be viewed with caution in Group 3 (the only experimental group to involve two written versions of the BDI-II). Because participants were given privacy to work on their packets, it is likely that a portion of the sample simply copied item responses from the first form, rather than read through the form a second time. However, there remains the argument that students possibly habituated to their environment, at least to some degree. An order effect, which was unexpected, was found.

Additional limitations of this study involve the imbalance in gender and ethnicity of the participants. There was not much variability in the sample. Women outnumbered men by a large margin, and the participants were predominantly Caucasian. Furthermore, given that the participants were recruited from psychology classes at a university, any generalizations beyond psychology undergraduate students, or to clinical populations, should be done cautiously.

Recommendations for Future Research

Despite these drawbacks, the study was successful in determining the measurement equivalence between the computerized and conventional forms of the BDI-II based on score comparisons. It should be noted that the establishment of measurement equivalence does not necessarily provide support for the equivalence of their construct validity. Given the dearth of studies available on the BDI-II, this study adds to the psychometric information on this instrument. The study also added data to the respective literature of the CAVS and the PMAQ, and generated a new questionnaire useful in quantifying computer experience (the CEI). Future research should include persons that may be most sensitive to constructs such as computer aversion, such as clinical populations, persons of different cultural backgrounds, the elderly, the poor, and those with disabilities (Nurius, 1990).

NOTES

Accepted for publication: July 18, 2000.

Portions of this article were presented at the 1999 meetings of the American Psychological Association, Boston. The authors thank Dr. Michael Granaas for statistical assistance with the Latin Squares analysis.

Address correspondence to Stefan Schulenberg, Doctoral Candidate, Department of Psychology, University of South Dakota, 414 E. Clark Street, Vermillion, South Dakota 57069. E-mail: SESchulen2@aol.com

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Beck, A. T., & Steer, R. A. (1993). *Manual for the Beck Depression Inventory.* San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment, 67,* 588–597.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8,* 77–100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Brown, D. T. (1984). Automated assessment systems in school and clinical psychology: Present status and future directions. *School Psychology Review, 13,* 455–460.

Butcher, J. N. (1987). The Use of Computers in Psychological Assessment: An Overview of Practices and Issues. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 3–14). New York: Basic Books, Inc.

Derogatis, L. R. (1983). *SCL-90–R administration, scoring, and procedures manual-II.* Towson, MD: Clinical Psychometric Research.

Dimock, P. H., & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement and Evaluation in Counseling and Development, 24,* 119–126.

Ford, B. D., Vitelli, R., & Stuckless, N. (1996). The effects of computer versus paper-and-pencil administration on measures of anger and revenge with an inmate population. *Computers in Human Behavior, 12,* 159–166.

Gardner, D. G., Discenza, R., & Dukes, R. L. (1993). The measurement of computer attitudes: An empirical comparison of available scales. *Journal of Educational Computing Research, 9,* 487–507.

George, C. E., Lankford, J. S., & Wilson, S. E. (1992). The effects of computerized versus paper-and-pencil administration on measures of negative affect. *Computers in Human Behavior, 8,* 203–209.

Harrell, T. H., Honaker, L. M., Hetu, M., & Oberwager, J. (1987). Computerized versus traditional administration of the Multidimensional Aptitude Battery-Verbal Scale: An examination of reliability and validity. *Computers in Human Behavior, 3,* 129–137.

Heinssen, R. K., Glass, C. R., & Knight, L. A. (1987). Assessing computer anxiety: Development and validation of the Computer Anxiety Rating Scale. *Computers in Human Behavior, 3,* 49–59.

Johnson, J. H. (1984). An overview of computerized testing. In M. D. Schwartz (Ed.), *Using computers in clinical practice: Psychotherapy and mental health applications* (pp. 131–133). New York, New York: The Haworth Press, Inc.

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.

Kjerulff, K. H., & Counte, M. A. (1984). Measuring attitudes toward computers: Two approaches. In G. S. Cohen (Ed.), *Proceedings of the Eighth Annual Symposium on Computer Applications in Medical Care* (pp. 529–535). New York: Institute of Electrical and Electronics Engineers.

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven's Standard Progressive Matrices, in particular for computerized testing. *European Review of Applied Psychology, 41,* 295–300.

LaLomia, M. J., & Sidowski, J. B. (1993). Measurements of computer anxiety: A review. *International Journal of Human-Computer Interaction, 5,* 239–266.

Lambert, M. E., & Lewis, D. (1989). *The Computer Experience Questionnaire. Unpublished manuscript.*

Meier, S. T. (1985). Computer aversion. *Computers in Human Behavior, 1,* 171–179.

Meier, S. T. (1988). Predicting individual differences in performance on computer-administered tests and tasks: Development of the Computer Aversion Scale. *Computers in Human Behavior, 4,* 175–187.

Meier, S. T., & Lambert, M. E. (1991). Psychometric properties and correlates of three computer aversion scales. *Behavior Research Methods, Instruments, & Computers, 23,* 9–15.

Merten, T. (1994). *Fragebogen zur Antwortbevorzugung.* Unpublished.

Merten, T., & Ruch, W. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences, 20,* 281–291.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22,* 71–83.

Nurius, P. S. (1990). A review of automated assessment. *Computers in Human Services, 6,* 265–281.

Olson-Buchanan, J. B., & Drasgow, F. (1999). Beyond Bells and Whistles: An Introduction to Computerized Assessment. In F. Drasgow & J.B. Olson-Buchanan (eds.), *Innovations in computerized assessment* (pp. 1–5). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

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.

Runyon, R. P., & Haber, A. (1991*). Fundamentals of behavioral statistics* (7th ed.). New York: McGraw-Hill, Inc.

Schulenberg, S. E., & Yutrzenka, B. A. (1999). The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. *Behavior Research Methods, Instruments, & Computers, 31,* 315–321.

Smith, R. E. (1963). Examination by computer. *Behavioral Science, 8,* 76–79.

SPSS (Version 7.5.1) [Computer software]. (1996). SPSS, Inc.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1997). Further evidence for the construct validity of the Beck Depression Inventory-II with psychiatric outpatients. *Psychological Reports, 80,* 443–446.

Steer, R. A., Rissmiller, D. J., Ranieri, W. F., & Beck, A. T. (1994). Use of the computer-administered Beck Depression Inventory and Hopelessness Scale with psychiatric inpatients. *Computers in Human Behavior, 10,* 223–229.

Tseng, H.-M., Macleod, H.A. & Wright, P. (1997). Computer anxiety and measurement of mood change. *Computers in Human Behavior, 13,* 305–316.

Turban, D. B., Sanders, P. A., Francis, D. J., & Osburn, H. G. (1989). Construct equivalence as an approach to replacing validated cognitive ability selection tests. *Journal of Applied Psychology, 74,* 62–71.