Measurement and Evaluation in
Counseling and Development

Examining Equivalency of the Driver Risk Inventory (DRI) Test Versions: Does it Matter Which Version I Use?

Journal:	*Measurement and Evaluation in Counseling and Development*
Manuscript ID:	MECD-14-01-003.R2
Manuscript Type:	Original Manuscript
Keywords:	equivalency testing, Driver Risk Inventory, computer-based testing
Abstract:	Equivalency of test versions is often assumed by counselors and evaluators. This study examined two versions, paper-pencil and computer-based, of the Driver Risk Inventory, a DUI/DWI risk assessment. An overview of computer-based testing and standards for equivalency provided. Results of the study confirmed reliability, validity and equivalency of the versions.

Examining Equivalency of the Driver Risk Inventory (DRI) Test Versions: Does it Matter Which Version I Use?

Abstract

Equivalency of test versions is often assumed by counselors and evaluators. This study examined two versions, paper-pencil and computer-based, of the Driver Risk Inventory, a DUI/DWI risk assessment. An overview of computer-based testing and standards for equivalency is also provided. Results of the study confirmed reliability, validity and equivalency of the versions.

Keywords: equivalency testing, Driver Risk Inventory (DRI) computer- based testing

Examining Equivalency of the Driver Risk Inventory (DRI) Test Versions: Does it Matter Which Version I Use?

Professionals have generally accepted computerized tests in counseling largely because of their long history of use in the field. This acceptance and familiarity has led to an expansion of computerized tests and scoring, albeit with concern that computerized score interpretations provided by test developers would replace professionals (Murphy & Davidshofer, 2001). There is scant evidence to support this concern. In fact, testing guidelines and recommendations advocate the use of collaborative evidence in conjunction with independent results (Greene, 2001; Murphy & Davidhofer, 2001). The American Counseling Association (2013) has established competencies for counselors who administer tests. Minimum competencies include sufficient training to appropriately select, administer, score and interpret test results. Moreover, the American Psychological Association (2013) requires that psychologists who administer tests have sufficient training to identify reporting errors and false positive results (American Psychological Association, 2013). Thus, as more computerized tests, modified versions, and imitations appear on the Internet, rather than becoming diminished, the role of qualified counselors and test administrators has expanded and the role of clinical judgment remains as great as ever.

Certainly the use of a modified or alternative versions is permissible and even appropriate under certain conditions; however, there must be supporting evidence that all versions are equivalent (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Counselors are doubtless very familiar with psychometric concepts like reliability and validity, but are largely unfamiliar with the concept of test equivalency--test equivalency is assumed, but rarely verified.

This paper explains equivalency testing of the Driver Risk Inventory (DRI), a DUI/DWI risk screening assessment. The DRI is widely used across North America to assess problems with alcohol, drugs, and driving behavior. The DRI can be administered using a paper-pencil format or online. This study examined the equivalency of the paper-pencil and online versions of the DRI.

Computer-based testing

There are two types of computerized testing commonly used: Computer- Based Testing (CBT) and Computer Adaptive Testing (CAT). The type of test, inventory or assessment usually determines which of these formats is preferable. As CBT is the format adopted by the DRI so this paper will limit its descriptions and explanations to this format.

Computer-based testing (CBT) is a modified version of a traditional/paper pencil (PP) test: test items are fixed, linear, and delivered in the same order. Software can be installed on a stand-alone computer, or the test can be accessed online. With CBT, scoring is done automatically and reports are typically generated within a few minutes. It is similar to paper-pencil administration in that it is highly structured--all items must be completed.

Researchers have argued that CBT offers some advantages over PP testing, including improved time efficiency, lower costs (Garb, 2007; Lewis et al., 2009), increased anonymity and confidentiality (Lewis et al.), minimization of cultural differences (Murphy & Davidhofer, 2001) and willingness of clients to disclose more with test administrators (Garb, 2007). Additionally, CBT measures features that cannot readily be assessed using paper-pencil measures, including response time, reading time and spatial abilities, as well as dynamic and patterned responses (Murphy & Davidhofer, 2001).

Paper-pencil testing does have its advantages, as it offers evaluator insights not captured by CB tests. Paper-pencil testing is generally considered more personal. Evaluators are able to observe client mood, affect, body language and interactions when using a PP format (Garb, 2007). Using a PP test also gives the counselor an opportunity to consider extenuating circumstances that may have impacted the testing situation or test score(s) (Butcher, Perry & Hahn, 2004). With valid PP versions, evaluators can be sure that the test measures the intended construct, and not test users' comfort or familiarity with technology.

Butcher and colleagues (2004) recommended that counselors and test administrators apply standards and guidelines established for PP testing to CB testing. The authors go on to suggest that, when possible, the test taking attitudes of test users be equivalent. They add that CB tests should be completed under controlled conditions, preferably in an office. In addition to similar external conditions, counselors should confirm that an alternative version demonstrates similar internal psychometric properties, including reliability and validity. As CBT usage expands, verification of test equivalency is essential.

Equivalency. Attempts to examine test equivalency have produced mixed results. Examination of alternative versions (PP versus CB) of the MMPI found no clinically relevant differences between test administration scores. When Fliege and colleagues (2009) examined the benefits of using CAT versus PP to assess depression in a group of patients, they found no clinically relevant differences between scores that could be attributed to the type of administration. It was noted that the PP administration provided insight into areas for item refinement that was lacking in the CAT administration. Fliege and colleagues also examined whether the role of use and comfort with technology contributed to depression test scores. They concluded it did not. Moreover, satisfaction surveys were administered to patients. The reported results indicated that the technology, a hand held device in this study, was generally well accepted by the patients.

Iverson et al., (2009) examined whether neuropsychological tests administered on the computer were as effective, accurate and valid as traditionally administered neuropsychological tests or whether computer familiarity was the construct being assessed. On some tests, specifically those tests that required rapid visual scanning and keyboard work, familiarity with computers did make a difference. Results suggested that the difference found on the CB version was enough to mimic a cognitive problem in some patients. Moreover, lack of computer familiarity can have a negative impact on computerized tests of reason. On the GRE subtests measuring analytical and quantitative domains, scores were negatively impacted on the CB version; however, on tests of language abilities and verbal intelligence, no meaningful difference between PP and CB scores was identified.

Early research and publications on computerized test scores, as summarized by Murphy and Davidhofer (2001), argued that score differences may be related to anxiety, response bias and faking good, novelty of using a computer, speed of information delivery, graphics, lengthy passages that a test taker must read on a screen, and the ability to omit or return to certain items. More recent research (Fliege, 2009; Iverson, Brooks, Ashton, Johnson, & Gualtieri 2009; Zitney et al., 2012) has explored these areas, but has also advocated increasing technology in testing (Greene, 2011). While appropriate for some tests (Zitny et al., 2012; Fliege, 2009), equivalency seems related to the construct under examination (Butcher et al., 2004), client or patient familiarity with computers (Iverson et al., 2009) and the complexity of the responses required (Fliege, 2009).

Methodology

Instrument

The DRI is a self-report assessment with 140 items that comprise five scales and a substance use classification. Items use true/false and multiple choice formats. The scales assess problems with alcohol and drugs, driver risk, stress management, and test truthfulness. In addition, the DRI uses a substance abuse classification that is derived from the DSM-IV. Items per scale are included in Table 4. The CBT version of the DRI uses the same test questions, and the items are presented in the same order. Software can be installed on a computer or the test can be accessed online.

A percentile score for the respondent's unique set of responses is generated for each scale and corresponds to the percentage of scores that fall below the given value in the frequency distribution of that scale. According to test developers, percentile scores between 0 and 39% represent a low risk; percentile scores between 40 to 69% represent a medium risk; scores between 70 and 89% represent a problem risk: and those with percentile scores between the 90th and 99th percentile are identified as having a severe problem (Behavior Data Systems, 2007). The substance abuse/dependency measure is based on DSM-IV classification criteria. The substance abuse/dependency classification is a binary measure of whether the respondent does or does not meet the substance abuse/dependency criteria.

The DRI has demonstrated concurrent validity (Chang, Gregory, & Lapham, 2002), the ability to distinguish between first time and multiple offenders (Leshowitz & Meyers, 1996) and the ability to identify problem drinkers (Jones & Lacey, 2000). DRI scales have demonstrated satisfactory reliability (α >.80) (Bishop, 2011a, Chang, 2002). Bishop (2011b) was able to demonstrate the predictive capabilities of the DRI for rapid DUI recidivist detection. Moreover, the National Transportation Highway and Safety Administration stated that the DRI is the only major DUI assessment that addresses driver risk (Popkins, 1988). Degiorgio and Lindeman (2013) found that Florida DUI recidivists demonstrated poorer stress management than offenders in the larger Florida DUI population, as measured by the DRI. Despite the substantial amount of research conducted on the DRI, no research has examined the equivalency of the traditional PP or CB versions of the DRI prior to this examination.

Participants

The study used data from Florida offenders who completed the DRI during 2012. The State of Florida mandates that all offenders complete the DRI regardless of whether they are convicted of a DUI. The choice of administering a DRI PP version or a CB version is determined by the agency overseeing the evaluation. Two samples were generated from the paper-pencil and online submissions. The PP sample consisted of 2, 520 offenders and the CB sample consisted of 2, 288 offenders. There were fewer PP submissions, so we oversampled this group to ensure adequate representation of offender characteristics. Demographic characteristics are presented in Table 1. As noted in the Table, there were differences in race/ethnicity and marital status between the two groups. Chi-square results revealed statistically significant group differences for race/ethnicity $x2$ (5) = 169.71, p <.001. Table 2 presents the offenders' self-reported arrests, driving infractions and blood alcohol concentration (BAC) at time of arrest. Reported arrests and infractions were similar for the groups; however, BAC averages were different, and were also found to be statistically significant t =6.03, $p >.001$.

Procedures

Three analyses were conducted to confirm the psychometric properties of the DRI versions. The primary analysis was a test equivalency study. Reliability analyses for the two versions were then conducted, followed by a validity study of each version.

Equivalency. Equivalency testing is relatively new to the social sciences, with its foundations in the field of bioresearch (Rogers, Howard, & Vessey, 1993). Equivalency can be measured by comparing means, dispersion and distribution shapes to determine whether the tests are similar. Several sources (Cribbie, Gruman, & Arpin-Cribbie, 2004; Lewis, Watson, & White, 2009; Mead & Drasgow, 1993; Rogers, Howard, & Vessey, 1993) indicate that hypothesis testing is considered inappropriate for equivalency testing because it suggests that there is insufficient evidence to reject the null (Lewis, et al., 2009). Cribbie and colleagues (2004) explain that the premise of the research question is fundamental; asking whether scores are different is not the same as asking whether the scores are similar. To address equivalency, the Schuirmann's approach described by Cribbe et al (2004) was applied. To that end a critical mean difference (D) was determined for each DRI scale. Raw scores were used because they generate the percentile rank and ultimately the offender's risk classification. A two-point difference became the level of confidence (Cribbie, 2004, p. 3) for the Truthfulness Scale, Alcohol Scale, Driver Risk Scale and Drug Scale. A 10-point difference was established for the Stress Management Scale. These differences were selected based on item scoring properties and their relationship to risk classifications. A mean difference that is smaller than the established critical mean would be considered unimportant and clinically irrelevant. Two one-sided hypotheses were used to establish equivalency, and one sided *t*-tests were conducted with a significance level of .05. Rejection of both hypotheses implies that the means are considered equivalent.

$$H_{01}: \ \mu_1 - \mu_2 > D; H_{02}: \mu_1 - \mu_2 > - D$$

Readers are directed to Cribbe et al. (2004) and Rogers et al. (1993) for more thorough explanations and examples.

Reliability. Cronbach's alpha was used to measure DRI item consistency. An analysis was conducted on each scale, for each version of the DRI. Table 4 lists the scales and summarizes the results for both groups of test takers.

Validity. Construct validity for each version was also examined using contrast groups. This approach differentiates offenders who are known to have higher risk factors from those known to have lower risk factors (DeVon et al., 2007). This study compared offenders' mean scale scores. Offenders were grouped into two categories based on the number of self-reported DUI arrests. Offenders with no more than one arrest were labeled first-time offenders and those with two or more arrests were labeled repeat offenders. It was anticipated that the repeat offenders would have higher mean scale scores than the first-time offenders on all scales except on the Stress Management Scale. On this scale higher scores are associated with better stress management skills. *T*-test analyses were conducted to examine whether the differences in mean scores were statistically significant. Adjusted *t* and *df* were used in the analysis along with a Bonferroni correction (p = .001) to control for experimentwise error (Field, 2009).

Results

Equivalency results indicate that PP and CB versions of the DRI were equivalent. A quick glance at Table 3 reveals that the raw score differences did not exceed the critical mean difference established *a priori*. In addition, statistical findings support the rejection of both null hypotheses for all DRI scales: the test means are considered equivalent. Results of the two independent *t*-tests are presented for review in Table 3.

Reliability coefficients for both groups were acceptable (α >.85). Validity results were consistent with the hypothesis: repeat offenders had higher mean scale scores on the Alcohol, Drug, and Driver Risk Scales. As expected, repeat offenders demonstrated poorer stress management (a lower mean score). First-time offenders had higher Truthfulness Scale scores that may be associated with their level of experience with law enforcement and assessment procedures. Repeat offenders' lower Truthfulness Scale scores may be attributed to awareness that denial, minimization, and deception would be detected. Table 5 presents the *t*-test results, as well as effect sizes. Cohen's *d* results found small to large effect sizes that were relatively consistent across the PP and CBT groups.

Early reviews of this manuscript pointed out that more sophisticated tests of invariance (equivalency) would address possible confounds related to differences in race/ethnicity. To that end, a post hoc multiple group confirmatory factor analysis (MGCFA) was conducted. MGCFA involves simultaneous CFA using groups for comparisons of population heterogeneity including mean structures, factor loadings, and variance (Brown, 2006). In this analysis CBT and PP groups were used and a test of *configural invariance* was conducted. According to Milfont & Fischer (2010), configural invariance measures whether the groups conceptualize the DRI constructs the same way. It is the least restrictive of the invariance tests but was appropriate to address the research question of this manuscript. Readers are directed to Vandenberg and Lance (2000) and Milfont and Fischer (2010) for comprehensive explanations of test invariance and specific model directions. Results demonstrated that, despite differences in race and ethnic composition, model fit was adequate; x^2 (16) = 241.35, p <.001; [LL 178.76; UL 279.38]; RMSEA .076.

Discussion

Computerized testing has a long history in the field of counseling, and its use among counselors is expected to grow. With the expansion, counselors must ensure that tests adapted or created specifically for computer administration have psychometric support and are equivalent to the traditional PP versions. This study examined the equivalency of the DRI PP and CB versions, as well as the psychometric properties of each version.

Test equivalency was established by comparing PP and CB mean scores through selecting a critical mean difference for each scale. The Truthfulness Scale, Alcohol Scale, Driver Risk Scale and Drug Scale had a 2-point critical mean difference. A 10-point critical mean difference was used for the Stress Management Scale. The use of raw scores and critical mean differences were selected based on the scoring properties of each scale. Two one-sided hypotheses were generated and statistical results confirmed that the PP version and CB versions of the DRI were equivalent. An ad hoc test of invariance revealed that both groups viewed the constructs underlying the DRI consistently.

Two additional analyses were conducted to confirm the psychometric properties of the PP and CB scores. Reliability scores of both DRI versions were satisfactory, with all coefficients greater than $\alpha > .85$. Construct validity was established through the use of contrast groups: offenders demonstrating greater risk had scores that reflected more problem severity. Results were consistent for both versions. These findings have added to the empirical support of the DRI as a DUI/DWI screening tool. The result of the equivalency analysis has also added empirical support for use of the CB version of the DRI. Counselors and evaluators currently using the DRI should have more confidence when administering either version of the DRI.

Despite the encouraging results there were some limitations that are worthy of mention. The groups used in the analyses were similar on several demographic and arrest-related variables, but there were some differences that should be noted. There were statistically significant differences between the two groups with regard to race/ethnicity and marital status. Also, while reported arrests and infractions were similar for the groups, BAC averages differed at statistically significant levels and may have resulted in offenders from the different sample groups being referred to different types of treatment/intervention programs. Additionally, it should be noted that this study used a sample of offenders from Florida and the results may not generalize to other States or other populations. Replications of this study using offenders from other populations and geographic locations are desirable and represent an area of future research.

In addition to offender characteristics, environmental influences may have played an important role in these analyses. As noted earlier, offender data were extracted from the Behavior Data Systems research database, which provides no information on the testing environment, setting or administration policies. Some researchers (Butcher et al., 2004) believe that the testing environment should be controlled to promote equivalency, and that should also apply to administration strategies and procedures. These criteria may be difficult to satisfy as CB testing limits administration to one test (unless the agency has multiple computers), while PP versions can easily be administered in large groups. Lewis et al (2009) found that an emotional response (laughter) in their large group administration may have influenced all test taker attitudes during the administration. Overt emotional responses, including anger or irritation are to be expected when screening offenders charged with criminal offenses. An offender exhibiting emotions during testing could influence all group members and their scores.

Another limitation which should be considered is item order and presentation. Computer-based administration of the DRI reveals one question at a time in a set format. In contrast to the PP version, where offenders can see all the items at one time and may answer in any order they choose. The difference in how items are presented may impact offenders' scores because it influences how an offender approaches and completes the test. It is important to note that neither item presentation approach completely protects against response sets.

In addition to replicating this study with other populations and geographic areas, collection of information about test administration procedures and environmental factors may provide insight into external factors that influence offenders' scores. Moreover, test equivalency of other inventories and assessments would also benefit from additional research and exploration.

Counselors, as well as test developers, have a responsibility to ensure that alternative versions of a test created and administered are reliable, valid and equivalent. A flawed test is flawed, no matter how it is administered (Garb, 2007). As computer-based testing expands, the role of the counselor will become more important in score interpretation and clinical assessment. The advantages of computer-based testing are many but should not substitute clinical judgment. As noted earlier, this is the first study to examine the equivalency of the DRI paper-pencil and computer-based test versions. Despite the limitations of this study, test administrators and counselors can have greater confidence that the DRI version they administer has empirical support and that the versions are equivalent.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington DC: American Education Research Association.

American Counseling Association (2013). Career Counseling Competencies. Retrieved from http://counseling.org/docs/competencies/career_counseling_competencies.pdf?sfvrsn=3

American Psychological Association. Testing Governance in APA. Retrieved from http://www.apa.org/science/programs/testing/governance.aspx

Ball, J. D., Archer, R. P., & Imhof, E. A. (1994). Time requirements of psychological testing: A survey of practitioners. *Journal of Personality Assessment, 63* (2), 239-249. doi:10.1002/jclp.10267

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: Guilford Press.

Butcher, J. N., Perry, J., & Hahn, J. (2004). Computers in clinical assessment: Historical developments, present status, and future challenges. *Journal of Clinical Counseling, 60* (3), 331-345.

Chang, I., Gregory, C., & Lapham, S.C. (2002). *Review of Screening Instruments and Procedures for Evaluating DWI Offenders.* Washington DC: AAA Foundation for Traffic Safety.

Creech, S. K., Evardone, M., Braswell, L., & Hopwood, C. J. (2010). Validity of the Personality Assessment Screener in veterans referred for psychological testing, *Military Counseling, 22,* 465-473. doi:10.1080/08995605.2010.513265

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Counseling, 60*(1), 1-10. doi: 10.1002/jclp.10217

Degiorgio, L., & Lindeman, H. (2013). Stress coping abilities and motivation for treatment among DUI recidivists. *Journal of Community Corrections, 12*(3) 5-9.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39* (2), 155-164.

Field, A. (2009). *Discovering Statistics Using SPSS.* (3rd Ed.). Washington DC: Sage.

Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. B., & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research, 18* (1), 23-26. doi:10.1002/mpr.274

Garb, H. N. (2007). Computer-administered interviews and rating scales. *Psychological Assessment, 19* (1), 4-13. doi: 10.1037/1040-3590.19.1.4

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology, 37,* 2, 261-278. doi:10.1111/j.1467-8535.2005.00525.x

Greene, R. L, (2011). Some considerations for enhancing psychological assessment. *Journal of Personality Assessment, 93,* 3, 198-203. doi:10.1080/00223891.2011.558879

Iverson, G. L., Brooks, B. L., Ashton, V. L., Johnson, L. G., & Gualtieri, C. T. (2009). Does familiarity with computers affect computerized neuropsychological test performance?

Journal of Clinical and Experimental Neurocounseling, 31, (5), 591-4-604. doi: 10.1080/13803390802372125

Lewis, I., Watson, B., & White, K.M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: equivalence testing of participants responses to health-related messages. *Australian Journal of Counseling, 61*(2), 107-116. doi:10.1080/00049530802105865

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 3, 449-458.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 2011-2084.

Racine, C. W., & Billick, S. B. (2012). Assessment instruments of decision-making capacity. *Journal of Psychiatry and Law, 40,* 243-263.

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

Vandenberg, R. J., & Lance, C. E. (2010). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(4), 4 -70. Doi: 10.1177/109442810031002

Zitny, P., Halama, P., Jelinek, M., & Kveton, P. (2012). Validity of cognitive ability tests-comparison of computerized adaptive testing with paper and pencil and computer-based forms of administration. *Studia Psychologica, 54*(3), 181-194.

Table 1

Participant Demographics

		Paper-pencil (N = 2530)	CBT (N = 2288)
Gender		%	%
	Male	69	73
	Female	30	27
Race/Ethnicity			
	White	83	68
	Black	4	9
	Hispanic	11	20
	Asian	<1	<1
	Native American	<1	<1
	Other	1	1
Marital			
	Single	50	58
	Married	21	20
	Divorced	21	17
	Separated	6	4
	Widowed	2	1
Education			
	8th Grade or less	4	2
	Some high school	11	12
	GED	7	7
	High school diploma	35	36
	Some college	2	2
	Technical/Business school	21	22
	College graduate	18	17
	Professional/graduate school	3	3

Table 2

Self-Reported Arrests and Driving Infractions

	CBT (N = 2288)				Paper-pencil (N = 2, 531)			
	Min	Max	Mean	SD	Min	Max	Mean	SD
DUI arrests	0	11	1.24	.82	0	10	1.35	.85
Reckless driving arrests	0	6	.20	.53	0	4	.16	.47
DUI arrests reduced	0	3	.16	.42	0	5	.14	.40
Alcohol-related arrests (not DUI)	0	30	.16	.82	0	5	.10	.40
Drug related arrests (not DUI)	0	6	.12	.42	0	10	.10	.42
At-fault accidents	0	12	.24	.59	0	5	.21	.53
Traffic tickets	0	20	.95	1.56	0	20	.75	1.37
Misdemeanors	0	20	.35	1.08	0	13	.35	.97
Felonies	0	30	.19	.93	0	30	.19	1.08
Arrests	0	4	1.10	.46	0	4	1.10	.45
BAC	.000	.397	.143	.08	.000	.390	.160	.07

Table 3

Equivalency results

Scales	PPT mean	CBT Mean	D	*t*1	*t*2	*p*
Truthfulness	10.68	12.12	1.44	-34.66	12.18	<.001
Alcohol	10.20	9.22	.98	-9.36	18.44	<.001
Drug	5.15	4.19	.89	-12.09	23.46	<.001
Driver Risk	9.76	8.87	.96	-12.41	22.88	<.001
Stress Management	137.50	143.82	-6.82	-16.58	3.74	<.001

Table 4

Reliability Coefficients

	Items	CBT	Paper-pencil
Truthfulness	21	.88	.88
Alcohol	23	.91	.92
Driver Risk	25	.85	.87
Drug	22	.91	.93
Stress Management	30	.90	.92

Table 5

Results for Construct Validity Using Contrast Groups

Paper-pencil test takers (N = 2,531)

Scales	First-time Offender	Multiple Offender	t	*p*	*d*
Truthfulness	10.97	9.96	3.54	<.001	.21
Alcohol	7.50	16.68	18.14	<.001	.96
Drug	4.37	6..89	6.22	<.001	.30
Driver Risk	8.55	12.53	10.72	<.001	.47
Stress Management	139.02	134.92	1.95	.05	.02
CBT test takers (N = 2, 288)					
Truthfulness	12.45	11.24	4.64	<.001	.17
Alcohol	6.82	15.45	17.20	<.001	.92
Drug	7.89	11.38	6.09	<.001	.29
Driver Risk	3.55	5.82	9.87	<.001	.48
Stress Management	144.21	143.23	.450	>.01	.09