Journal of the International Neuropsychological Society (2007), **13**, 799–806.

DOI: 10.1017/S1355617707071020

The Indiana University Telephone-Based Assessment of Neuropsychological Status: A new method for large scale neuropsychological assessment

FREDERICK W. UNVERZAGT,[1] PATRICK O. MONAHAN,[2] LYNDSI R. MOSER,[1] QIANQIAN ZHAO,[2] JANET S. CARPENTER,[3] GEORGE W. SLEDGE, JR.,[2] AND VICTORIA L. CHAMPION[3]

[1]Department of Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana
[2]Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana
[3]Indiana University School of Nursing, Indianapolis, Indiana

(RECEIVED October 27, 2006; FINAL REVISION March 6, 2007; ACCEPTED March 8, 2007)

Abstract

Sensitive measures of neuropsychological function were adapted to a telephone administration format for use in a large survey of quality of life in breast cancer survivors (BCS). Healthy controls (HC) and BCS were recruited from the community and administered the same neuropsychological test battery on two occasions separated by 1 week. Subjects were randomly assigned to conditions, stratified by diagnosis: In-person at Time-1 and In-person at Time-2 (P-P); Telephone at Time-1 and Telephone at Time-2 (T-T); T-P; and P-T. Four cognitive (Rey Auditory Verbal Learning Test, Controlled Oral Word Association, Digit Span, Symbol Digit) and two self-report measures (Squire Memory Self-Report Scale, Center for Epidemiological Studies Depression Scale) were used. The 106 subjects were randomized (54 HC and 52 BCS). Test–retest reliabilities (intraclass correlations) did not differ significantly by condition across the cognitive or self-report measures and ranged from moderate to near perfect (r's .43–.93; p's $< .05$). Mean scores at Time-1, practice effects (Time-1 to Time-2), and standard errors of measurement were comparable between In-person and Telephone administration formats. Results suggest that memory, attention, information processing speed, verbal fluency, and self-report of mood and memory can be measured reliably and precisely over the telephone. (*JINS*, 2007, *13*, 799–806.)

Keywords: Mass screening, Neuropsychological test, Cognition, Memory, Mild cognitive Impairment, Diagnosis, Dementia

INTRODUCTION

Up to 30–50% of breast cancer survivors report persistent problems with memory and concentration (Berglund et al., 1991; Hurria et al., 2006). Treatment-related cognitive dysfunction among breast cancer survivors has also been identified in studies using objective measurements of cognitive functioning (Ahles et al., 2002; Brezden et al., 2000; Tchen et al., 2003; van Dam et al., 1998; Wefel et al., 2004). A recent meta-analysis of the relationship between cognitive dysfunction and adjuvant chemotherapy for breast cancer found the largest effect sizes for measures of memory (including list learning tests) and language [including verbal fluency tests (Stewart et al., 2006)]. One factor that has hampered research in this area is that most neuropsychological testing is done in-person in a laboratory and can require hours to complete. There is a need for an assessment of cognitive function using sensitive tests that can be accomplished without the lengthy in-person methodology.

Telephone-based assessment has the advantage of allowing large scale assessment as is required in epidemiological and survey research; however, currently available telephone-based cognitive assessments are based on mental status tests, for example, Telephone Interview for Cognitive Status (TICS, [Brandt et al., 1988]) and TICS-modified (Plassman et al., 1994; Welsh et al., 1993), which are known to have limited sensitivity to mild cognitive dysfunction (Tombaugh & McIntyre, 1992). On the other hand, list learning,

Correspondence and reprint requests to: Frederick W. Unverzagt, Ph.D., Department of Psychiatry, Indiana University School of Medicine, 1111 W. 10th Street, Suite PB 218A, Indianapolis, IN 46202, USA. E-mail: funverza@iupui.edu

verbal fluency, information processing speed, and digit repetition are sensitive to cognitive dysfunction across a wide range of neurologic and psychiatric conditions (Christensen et al., 1991; Zakzanis et al., 1999). It is precisely these more sensitive tests that are needed to accurately measure cognitive deficits that may be associated with breast cancer and its treatment.

Our purpose was to adapt sensitive measures of memory, attention, information processing speed, verbal fluency, and mood to the telephone administration format so they could be used in a large survey of quality of life in breast cancer survivors. In this study, we examined the psychometric properties of a short battery of neuropsychological tests as a function of method of administration: standard in-person *versus* a telephone administration format. We hypothesized that (a) test–retest reliabilities would be comparable as a function of method of administration and similar to published standards, (b) measurement precision would be comparable across method of administration (i.e., similar standard errors of measurement), (c) practice effects would be comparable as a function of method of administration (i.e., no significant differences between in-person and telephone administration formats with regard to change in scores over time), and (d) initial cognitive test performance would be comparable across method of administration.

METHOD

Sampling Frame

This study was approved by and subject to ongoing review by the Institutional Review Board of Indiana University–Purdue University Indianapolis. Participants were female breast cancer survivors (BCS) and healthy controls (HC). Breast cancer survivors were recruited from a cancer research registry, cancer support groups, nominations of enrolled subjects, and advertisements posted at local churches and community centers. Eligibility was determined in an initial telephone screening interview. Verbal consent was obtained at the outset of the screening. Inclusion criteria for BCS subjects were (1) self-reported history of breast cancer, (2) at least 1 year after completion of local breast cancer treatment (chemotherapy, radiation, surgery), (3) living independently in the community, (4) absence of self-reported major psychiatric disorder (major depression, bipolar disorder, history of schizophrenia, or psychosis from any cause) or neurologic condition (learning disability, head injury with loss of consciousness greater than 60 min, epilepsy, stroke, brain tumor, brain infection, or brain degeneration), and (5) 40 years of age and older. Participants with a history of metastatic cancer, recurrent cancer, or other cancers, with the exception of skin cancer, were excluded.

The HC were recruited from a research registry, nominations of enrolled subjects, and advertisements posted at local churches and community centers. Eligibility was determined in an initial telephone screening interview. The criteria were identical to those for the BCS except that self-reported history of any cancer (other than skin cancer) had to be absent and that only HC that individually matched in age (± 5 years) and education (± 3 years) to an enrolled BCS were eligible. The selection criteria were designed to result in a mixed sample of BCS and HC subjects that would match the general demographic characteristics of the Eastern Cooperative Oncology Group (ECOG), Quality of Life Study cohort, from the American Cancer Society grant (RSGPB-04-089-01-PBP).

Design

Upon completion of the eligibility screening, participants were scheduled for an appointment at the research center. Upon arrival at the center, participants gave written informed consent and were randomly assigned to one of four conditions, stratified by subject status (BCS *vs*. HC). Each subject accepted into the study was administered the same neuropsychological test battery on two occasions separated by 1 week. The four conditions were (1) In-person at Time-1 and In-person at Time-2 (P-P), (2) Telephone at Time-1 and Telephone at Time-2 (T-T), (3) Telephone at Time-1 and In-person at Time-2 (T-P), and (4) In-person at Time-1 and Telephone at Time-2 (P-T). Participants were paid $25 for each appointment ($50 total). The neuropsychological test battery was individually administered by trained and experienced psychometricians in an office in the research center. When a telephone administration format was used, the subject was placed in an office in the research center and the psychometrician called the subject over a hard-wired telephone line using standard business-grade telephones. Study design is depicted in Figure 1.

Test Battery

The neuropsychological battery consists of standard clinical instruments measuring new learning and recall, attention, information processing speed, verbal fluency, and self-reported memory and mood that have been in wide clinical use for many years (Lezak et al., 2004). Tests are listed in the order administered: Rey Auditory Verbal Learning Test (AVLT; Rey, 1941), total learning is the sum of words from the five learning trials; WAIS-III Digit Span (Wechsler, 1997); Symbol Digit Modalities (Smith, 1982), oral response format; Controlled Oral Word Association (COWA; Benton & Hamsher, 1989), for which published procedures were augmented so that whenever a subject gave a response beginning with a letter other than the target letter, the examiner re-established the correct target letter by providing these phonetic examples: “No, say words that start with the letter ‘C’ as in ‘Charlie’ (‘F’ as in ‘Fred’ or ‘L’ as in ‘Linda’)”; Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977), a self-report questionnaire for depression with higher scores indicating more depression; and Squire Memory Self-Report Scale (SRS; Squire & Zouzounis, 1988), an 18-item self-

https://doi.org/10.1017/S1355617707071020 Published online by Cambridge University Press

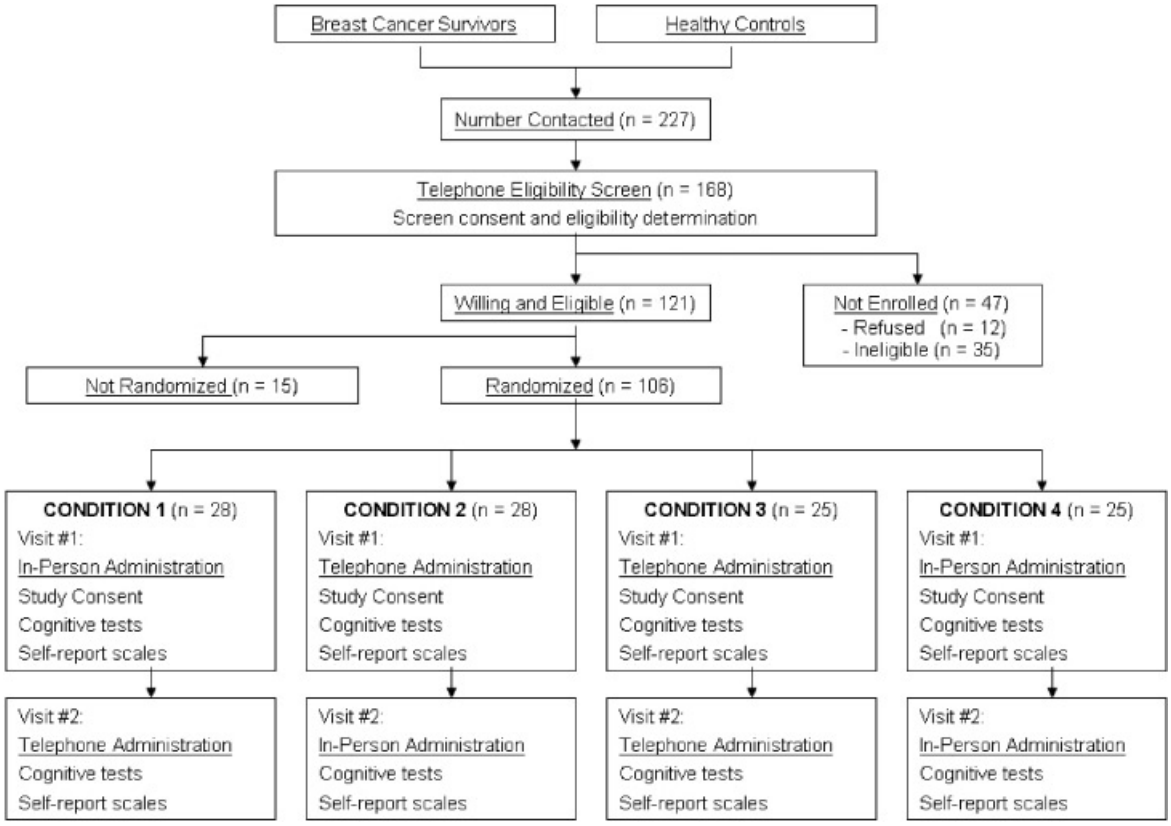

Fig. 1. Study design.

assessment of memory compared with the average person on a 9-point scale (1 = "Worse than the Average Person," 5 = "Same as the average person," and 9 = "Better than the Average Person"). AVLT long delayed recall and recognition were taken after the SRS (between 15 and 35 min after the last immediate recall trial).

Adaptations for Telephone-Based Assessment

Before the telephone assessment, subjects were given stimulus sheets (Symbol Digit form and self-report response options) and instructed to keep these sheets nearby during the call. They were told not to mark or write on the forms at any time.

At the conclusion of each assessment, the examiner made ratings of the subject's hearing, comprehension, and behavioral and attitudinal response to testing on Likert scales. Each assessment was also coded for validity on a 3-point scale (adequate, borderline, inadequate) using clinical judgment based on quality of the examination (e.g., telephone connection, extraneous noises, interruptions) and subject characteristics (e.g., hearing, confusion, abnormal motivation or effort, and abnormal events such as sounds of writing or perfect serial order responding on word list recall). Motivation and effort were rated according to the subject's general cooperativeness, which is captured, in part, by how readily the subject abandons tasks and expresses negative emotions (e.g., frustration, anger, hopelessness).

Statistical Analysis

The demographic characteristics of the subjects across method of administration at Time-1 were compared using a t test (for continuous data) and a χ^2 test (for categorical data). Statistical analysis of hypothesis "a" consisted of estimating reliability with the intraclass correlation coefficient (ICC) derived from a two-way random effects model using the consistency definition (McGraw & Wong, 1996). The consistency definition was important to use in this study because a constant true score improvement for all persons, due to practice effects, are likely to occur with cognitive exams from test to retest. With a sample size of at least 25 persons in each condition, this study is powered to provide 95% confidence intervals (CI) with a distance of .20 or less on each side of ICC reliability estimates. The ICC for the T-T and P-P conditions represented test–retest reliability coefficients. The ICC for the T-P and P-T conditions represented a combination of interprocedure and test–retest reliability, which provides a conservative estimate of interprocedure reliability. To compare the ICC across the four conditions, we performed an approximate test by examining whether their 95% CI overlapped. Two ICC whose CI did not overlap would be considered statis-

https://doi.org/10.1017/S1355617707071020 Published online by Cambridge University Press

tically significantly different. However, another important aspect of hypothesis "a" was to descriptively describe the ICC in relation to published norms.

To examine precision (hypothesis "b"), standard errors of measurement (*SEM*) were calculated using condition-specific test–retest values and the average of the Time-1 and Time-2 *SD*s for each test. These *SEM* were analyzed descriptively. In addition, for hypothesis "c," practice effects (difference scores from Time-2 to Time-1) were tested separately for P-P and T-T conditions using the two-sided paired *t* test; then, the comparability of practice effects between method of administration was tested by using the two-sided independent samples *t* test for comparing the mean change scores for P-P *versus* T-T conditions. We examined the effect of method of administration on performance (hypothesis "d") by conducting independent samples *t* tests on each measure at Time-1, combining conditions with similar administration formats at Time-1 [P-T combined with P-P (i.e., P at Time-1), and T-P combined with T-T (i.e., T at Time-1)].

Table 1. Sample characteristics by method of administration at Time-1

	In-person administration (n = 53)		Telephone administration (n = 53)		
	Count	%	Count	%	p value
Age, years[a]	58.9	8.9	58.6	9.3	.86
Education, years[a]	15.0	2.7	14.9	2.5	.88
Race					1.0
White	45	84.9	45	84.9	
Non-white	8	15.1	8	15.1	
Marital status					.31
Married	31	58.5	36	67.9	
Nonmarried	22	41.5	17	32.1	
Diagnostic group					1.0
Healthy Control	27	50.9	27	50.9	
Breast Cancer Survivor	26	49.1	26	49.1	
Hearing loss					.46
Yes	12	22.6	9	17.0	
No	41	77.4	44	83.0	

[a]Values are mean and *SD*.

RESULTS

A total of 227 women were contacted to participate in the study, and 168 (74%) completed the initial screening. Of the 168 women who completed the screening, 121 (72%) were enrolled and 47 were not (35 ineligible, 12 refused). A total of 106 subjects were randomized (15 of the enrolled subjects did not attend the first assessment session and were not randomized). Almost 85% of the randomized subjects were white [the rest were African American (14.2%) or biracial (0.9%)], average age was 58.8 (± 9.0) years, mean education was 15.0 (± 2.6) years, and just over 63% were married [the rest were divorced (15.1%), widowed (14.2%), never married (5.7%), or living as married (1.9%)]. Of the participants, 54 were HC and 52 were BCS. Twenty-one participants self-reported hearing loss but only two used hearing aides.

The 106 randomized participants and the 27 subjects that were eligible but not randomized (the 12 who refused plus 15 who were not randomized) did not differ significantly in education [$t(131) = .99$; $p = .32$], proportion of whites [$\chi^2(133) = .001$; $p = .97$], or proportion of married [$\chi^2(133) = .112$; $p = .74$].

In Table 1, the sample has been collapsed across mode of administration (In-person *vs*. Telephone) at Time-1. There were 53 participants in each format. There were no statistically significant group differences in age [$t(104) = .18$; $p = .86$], education [$t(104) = .15$; $p = .88$], proportion of whites [$\chi^2(106) = .0$; $p = 1.0$], proportion of married [$\chi^2(106) = 1.01$; $p = .31$], proportion of HC [$\chi^2(106) = .0$; $p = 1.0$], or proportion with self-reported hearing loss [$\chi^2(106) = .53$; $p = .46$].

The test–retest and interprocedure plus test–retest reliability coefficients for the cognitive and self-report measures are reported in Table 2 by condition. The estimates for the P-P condition are all statistically significant. The estimates for the cognitive scores range from .62 (AVLT long delay) to .90 (COWA) and fall in the range of "substantial" to "almost perfect" (terms from Landis & Koch, 1977). The reliability of self-reported depression was "moderate" (CES-D $r = .43$; $p < .05$), while self-ratings of memory were "almost perfect" (Squire SRS $r = .87$; $p < .001$). In the T-T condition, all ICCs are all statistically significant and in the "substantial" to "almost perfect" range (from .73 on CES-D to .93 on Squire SRS). The 95% confidence intervals for each estimate overlap between conditions; therefore, there are no significant differences in test–retest reliability as a function of method of administration. The test–retest reliabilities in the T-P and P-T conditions follow the same patterns, suggesting that crossing the method of administration in the test–retest format results in scores that are "substantial" to "almost perfect" in their correspondence. All sets of estimates very closely parallel the test–retest reliabilities reported in other published studies and test manuals.

Table 3 presents information on practice effects and precision (*SEM*) as a function of method of administration. Practice effects were calculated as the difference between Time-2 and Time-1 mean scores for each test within P-P and T-T conditions. Precision of measurement was calculated as *SEM* for each test under each condition (using the condition-specific r_{tt} and *SD*). For both P-P and T-T conditions, there were significant practice effects (improvement) in scores from Time-1 to Time-2 for AVLT total learning, AVLT long delay, Symbol Digit number-correct, and COWA total score (paired samples *t* tests, all *p*'s ≤ .05). There were no significant changes in mean scores from Time-1 to Time-2 for Digit Span total score, CES-D, or Squire SRS total score, for either T-T or P-P (paired samples *t* tests, all *p*'s > .05). The practice effects were not statistically different between P-P and T-T conditions (see rightmost column in Table 3);

https://doi.org/10.1017/S1355617707071020 Published online by Cambridge University Press

Table 2. Reliability coefficients by condition

	Condition				
	P-P ($n = 25$)	T-T ($n = 25$)	T-P ($n = 28$)	P-T ($n = 28$)	
Test score	ICC (test–retest)	ICC (test–retest)	ICC (method & test–retest)	ICC (method & test–retest)	Published r_{tt}
AVLT Total learning	.84***	.78***	.82***	.79***	.77[a]
Long delay	.62***	.82***	.47**	.85***	.60[b]
Digit Span, total	.78***	.75***	.82***	.81***	.85[c]
Symbol Digit, number correct	.81***	.86***	.87***	.91***	.76[d]
COWA, total score	.90***	.88***	.90***	.78***	.70[e]
CES-D, total score	.43*	.73***	.65***	.88***	.75[f]
Squire SRS, total score	.87***	.93***	.90***	.92***	–

Note. P-P = Person-Person condition (see text for description); T-T = Telephone-Telephone; T-P = Telephone-Person; P-T = Person-Telephone; AVLT = Auditory Verbal Learning Test; COWA = Controlled Oral Word Association; CES-D = Center for Epidemiological Studies-Depression Scale; SRS = Squire Subjective Memory Questionnaire.
[a]Geffen et al., 1994.
[b]Uchiyama et al., 1995.
[c]Tulsky et al., 1997.
[d]Smith, 1982.
[e]Ross, 2003.
[f]Maloni et al., 2005.
*$p < .05$.
**$p < .005$.
***$p < .001$.

that is, the independent samples *t* test comparing the two groups (P-P *vs*. T-T) on change scores found no significant differences, indicating that these methodologies return comparable mean scores over time. Additionally, the *SEM*s based on test–retest reliability were descriptively very similar for the two groups, suggesting comparable measurement precision for the two methods.

To examine the effect of method of administration (In-person *vs*. Telephone) on mean performance, the elapsed times, cognitive scores, and self-reported mood and memory ratings are presented by condition collapsed at Time-1 in Table 4. There were no significant differences in scores between In-person and Telephone administration on any cognitive measure. For the major cognitive indices, the difference scores between In-person and Telephone administration were very small. Using Cohen's descriptors (Cohen, 1988) for effect sizes ($d = [M_{Person} - M_{Telephone}]/SD_{Pooled}$), the effect of administration was small for all cognitive indices (e.g., AVLT total learning d = .19, long delayed recall d = .11).

Table 3. Practice effects and *SEM* by test score by condition

	P-P ($n = 25$)		T-T ($n = 25$)		
	Practice Effect	*SEM*	Practice Effect	*SEM*	Between groups *p* value[a]
AVLT Total learning	9.96*	3.40	10.12*	3.69	.91
Long delay	1.84*	1.69	2.36*	1.07	.36
Digit Span, total	.56	2.08	1.08*	1.82	.51
Symbol Digit, number correct	3.20*	3.56	3.44*	3.16	.86
COWA, total score	2.04*	3.14	3.44*	3.94	.33
CES-D, total score	−1.84	4.20	−2.20	4.32	.86
Squire SRS, total score	1.64	8.93	−.28	4.71	.50

Note. P-P = Person-Person condition (see text for description); T-T = Telephone-Telephone condition; Practice effect = $Mean_{Time\text{-}2} - Mean_{Time\text{-}1}$; *SEM* = standard error of measurement [$SD\sqrt{(1 - r_{tt})}$ with *SD* the average standard deviation of Time-1 and Time-2 and r_{tt} of Time-1 × Time-2]; AVLT = Auditory Verbal Learning Test; COWA = Controlled Oral Word Association; CES-D = Center for Epidemiological Studies-Depression Scale; SRS = Squire Subjective Memory Questionnaire.
[a]Comparison of differences in practice effects between conditions.
*$p \leq .05$ for the difference between Time-1 and Time-2 score within condition.

https://doi.org/10.1017/S1355617707071020 Published online by Cambridge University Press

Table 4. Mean test scores by mode of administration at Time-1

	In-person administration (n = 53)		Telephone administration (n = 53)		Difference between methods (In-person—Telephone)	
	Mean	*SD*	*Mean*	*SD*	Mean raw score	Effect size (d)
AVLT Total learning (max = 75)	51.0	8.2	49.5	7.5	1.5	.19
Total confabulations	1.4	3.4	1.3	1.7	.1	.04
Trial B (max = 15)	5.6	1.5	5.7	1.7	−.1	−.06
Short delay (max = 15)	10.1	2.9	9.8	3.0	.3	.10
Long delay (max = 15)	10.4	2.7	10.1	2.9	.3	.11
Delay confabulations	.4	.7	.4	.7	.0	.00
Delay intrusions	.2	.4	.2	.5	.0	.00
Recognition A (max = 15)	14.0	1.4	13.8	1.6	.2	.13
Recognition B (max = 15)	14.1	1.9	14.1	1.6	.0	.00
False positives	2.2	4.1	2.0	2.6	.2	.06
Digit Span, forward (max = 16)	10.1	2.6	10.8	2.3	−.7	−.29
Digit Span, backward (max = 14)	7.2	2.2	7.7	2.1	−.5	−.23
Digit Span, total (max = 30)	17.3	4.4	18.5	3.7	−1.2	−.30
Symbol Digit, correct (max = 110)	52.8	9.3	55.0	9.4	−2.2	−.24
Symbol Digit, errors	1.1	1.5	.7	1.2	.4	.29
COWA, total score	39.8	10.8	41.1	12.9	−1.3	−.11
CES-D, total score (max = 60)	8.3	6.4	11.6	9.3	−3.3*	−.41
Squire SRS, total score (max = 162)	97.7	22.5	98.0	19.1	−.3	−.01
Elapsed time, minutes	36.0	5.0	33.0	4.0	3.0*	.66

Note. AVLT = Auditory Verbal Learning Test; COWA = Controlled Oral Word Association; CES-D = Center for Epidemiological Studies-Depression Scale; SRS = Squire Subjective Memory Questionnaire; ns = not significant.
*$p \leq .05$.

The Telephone format was associated with a significantly shorter elapsed time for administration (33.0 ± 4.0 min) compared with In-person [36.0 ± 5.0 min; $t(104) = 2.12$; $p = .04$]. The Telephone format was also associated with a significantly greater report of depressive symptoms on the CES-D (11.6 ± 9.3) compared with the In-person format [8.3 ± 6.4; $t(104) = 2.12$; $p = .04$]. We examined whether inattentiveness in the telephone assessment could have resulted in these higher CES-D scores. The four reverse-keyed items on the CES-D (e.g., "I felt hopeful about the future") were grouped to form a subscale, and the 16 normal-keyed items were grouped to form a subscale (e.g., "I felt depressed"). For each subject, a difference score was calculated between these two subscales. Nonattentive subjects might be expected to fail to shift out of the predominant response mode causing the subscale difference score to be closer to zero, while attentive subjects might be expected to have scores further from zero. Scores differed significantly by administration format [$t(104) = 2.29$; $p < .03$]. The Telephone group ($M = -8.49$; $SD = 7.2$) was further from zero than the In-person group ($M = -5.90$; $SD = 4.0$), suggesting that the participants in the Telephone group were more attentive to shifts in item content.

A small number of subjects self-reported hearing loss ($n = 21$). A general linear model with self-reported hearing loss (yes *vs*. no) and method of administration (In-person *vs*. Telephone) as fixed factors and age and education as covariates on Time-1 scores was conducted. Results revealed no main effects for hearing and no interaction of hearing with method of administration on any of the main cognitive and self-report indices [all $F(5,100) \leq 1.5$; $p > .23$].

DISCUSSION

In this mixed sample of healthy controls and breast cancer survivors, we have demonstrated that a telephone administration format captures cognitive test scores and self-reported mood and memory ratings just as reliably and precisely as the traditional in-person method. The test–retest correlations, standard error of measurement, practice effects, and mean scores for the main performance measures from tests of new learning, working memory, information processing speed, and verbal fluency and self-report measures of mood and memory were similar whether obtained *via* in-person examination or over the telephone. Test–retest correlations for the AVLT, Digit Span, Symbol Digit, and COWA in the T-T condition were "substantial" to "almost perfect" (Landis & Koch, 1977) and very comparable to retest correlations reported in the literature using these same tests in an in-person format (Geffen et al., 1994; Maloni et al., 2005; Ross, 2003; Smith, 1982; Tulsky et al., 1997; Uchiyama et al., 1995). Somewhat lower reliability estimates were observed for AVLT delayed recall and self-report of depression on the CES-D (possibly due to small interquartile range on the raw scores). In general, our test–retest reliability estimates are comparable to results obtained

https://doi.org/10.1017/S1355617707071020 Published online by Cambridge University Press

by Plassman and colleagues on the TICSm in a mixed sample of 67 healthy controls and cognitively impaired subjects (r_{tt} = .83; Plassman et al., 1994) and slightly below the r_{tt} = .97 found by Brandt et al. on the TICS in a small sample of 34 dementia patients (Brandt et al., 1988).

There were two areas where a significant difference occurred related to format of administration. First, total test times were slightly *shorter* for the telephone approach. Whereas the 3-min time savings in favor of telephone administration is small and unlikely by itself to be a reason to choose this method of administration over an in-person method, it was nonetheless an unexpected finding. It may be that less "small talk" took place over the telephone, perhaps due to the reduced strength of the interpersonal connection in that format *versus* being in the physical presence of the examiner.

Second, the self-ratings of depression were actually *higher* in the telephone-based format than in the in-person format. Again, it may be that the absence of the examiner's physical presence in the telephone format created a slightly greater perception of anonymity, which decreased the demand for socially desirable responding. These data are consistent with other studies showing freer self-disclosure in comparisons of Web-based *versus* traditional survey methods (Parks et al., 2006). Analysis of difference scores between reverse- and normal-keyed items on the CES-D indicated that differences in attentiveness do not underlie the mode-of-administration effect on self-report depression scores. It appears that the reduced interpersonal demands of the telephone format both hastened the assessment and may have allowed for more veridical self-report of affect.

Our study extends previous research in this area, which has focused on relatively less sensitive dementia screening tools such as the TICS (Brandt et al., 1988; Plassman et al., 1994; Welsh et al., 1993). The tests in the Indiana University Telephone-Based Assessment of Neuropsychological Status (IU-TBANS) are among the most sensitive to brain dysfunction across a wide range of neurologic and psychiatric conditions (Zakzanis et al., 1999). Although our results need to be considered in light of our sample (i.e., relatively healthy, well-educated, middle-aged women), the extension of telephone-based assessment to sensitive measures of brain function improves the options available for large scale neuropsychological assessment in survey and epidemiological research.

Telephone-based assessment has distinct advantages for the subject by saving time and effort associated with physical travel to and from clinic and allowing greater flexibility in scheduling the time that assessment occurs. These features are likely to combine to improve overall response rates. The benefits of telephone-based cognitive assessment to the researcher include improved time efficiency by eliminating travel by staff, improved safety by eliminating home visits, and conserved physical resources by eliminating the need for clinic space to obtain data. These features make it possible to conduct large scale assessments as occurs in epidemiological and survey research.

Telephone-based assessment in the field also has distinct challenges. Our study was conducted in a laboratory using hard-wired telephone connections. Our results may not generalize to actual testing conditions in the field where cellular and wireless telephones may produce reduced audio fidelity and delays in transmission that could affect the assessment. In our field work, we give preassessment suggestions to help structure the assessment, including (1) to be sure the room was quiet and free from distractions and interruptions; (2) to turn off any radios, TVs, or cell phones; (3) to clear all papers, magazines, books, and writing utensils from the vicinity; (4) to have the stimulus sheets within arm's reach and placed face down; (5) to be prepared to ignore any call-waiting interruptions; and (6) to not write any notes during the session. This type of optimal structure may not always be possible in the field, and performance could be negatively affected as a result.

We had a small number of subjects with self-reported hearing loss. Although they performed as well as normally hearing subjects on the telephone assessment, subjects in the field with more severe hearing loss and those with hearing aides may have difficulty understanding instructions and responding over the telephone.

Another challenge associated with the telephone methodology is the lack of a visual channel. Detailed and sensitive assessment of spatial processing may not be possible by this method. The lack of a visual channel also makes it hard to monitor nonstandard behavior (i.e., some forms of cheating). While our examiners make behavioral and validity ratings based the quality of the assessment environment and subject factors, and these ratings help to allow tracking of the quality of the data, they do not substitute for the tight control of the environment and close visual observation of the subject that is afforded by in-person assessments in the laboratory.

The modest sample size in this study means that the study is not powered to detect small changes and the risk of type II error in interpreting nonsignificant differences exists. The absolute differences between cognitive scores based on method of administration are small—less than 3 raw score points in all cases and frequently a fraction of a raw score point and the effect sizes are all small in magnitude, less than or equal to .30 in all cases (Cohen, 1988). While the risk of a type II error exists, the practical significance of the difference is likely small. It is also true that the majority of the test-retest reliabilities across the four conditions were greater than .80 and that only 4 of 28 re-test reliability estimates were less than .70. Overall, it would appear that telephone administration of these tests and measures has reliability in conventionally acceptable ranges. As always, application of these findings to individuals that are very different from this sample should be done cautiously if at all. In conclusion, this study has established that the IU-TBANS is a reliable and precise telephone-based neuropsychological assessment composed of sensitive tests of new learning, attention, information processing speed, executive function, and mood that may facilitate large scale

neuropsychological assessment in epidemiological and survey research.

ACKNOWLEDGMENTS

We thank Ms. Anne Murphy-Knudsen and Ms. Sara Hickey for their assistance. This work is supported by grants from the American Cancer Society (RSGPB-04-089-01-PBP), the Mary Margaret Walther Program of the Walther Cancer Institute (100-200-20572), and the National Institute on Aging (P30 AG10133).

REFERENCES

Ahles, T.A., Saykin, A.J., Furstenberg, C.T., Cole, B., Mott, L.A., Skalla, K., Whedon, M.B., Bivens, S., Mitchell, T., Greenberg, E.R., & Silberfarb, P.M. (2002). Neuropsychologic impact of standard-dose systemic chemotherapy in long-term survivors of breast cancer and lymphoma. *Journal of Clinical Oncology*, *20*, 485–493.

Benton, A.L. & Hamsher, K.D. (1989). *Multilingual Aphasia Examination*. Iowa City, Iowa: AJA Associates.

Berglund, G., Bolund, C., Fornander, T., Rutqvist, L.E., & Sjoden, P.O. (1991). Late effects of adjuvant chemotherapy and postoperative radiotherapy on quality-of-life among breast-cancer patients. *European Journal of Cancer*, *27*, 1075–1081.

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *1*, 111–117.

Brezden, C., Phillips, K., Abdolell, M., Bunston, T., & Tannock, I. (2000). Cognitive function in breast cancer patients receiving adjuvant chemotherapy. *Journal of Clinical Oncology*, *18*, 2695–2701.

Christensen, H., Hadzi-Pavlovic, D., & Jacomb, P. (1991). The psychometric differentiation of dementia from normal aging: A meta-analysis. *Psychological Assessment*, *3*, 147–155.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Geffen, G.M., Butterworth, P., & Geffen, L.B. (1994). Test-retest reliability of a new form of the Auditory Verbal Learning Test (AVLT). *Archives of Clinical Neuropsychology*, *9*, 303–316.

Hurria, A., Goldfarb, S., Rosen, C., Holland, J., Zuckerman, E., Lachs, M.S., Witmer, M., Van Gorp, W.G., Fornier, M., D'Andrea, G., Moasser, M., Dang, C., Van Poznak, C., Robson, M., Currie, V.E., Theodoulou, M., Norton, L., & Hudis, C. (2006). Effect of adjuvant breast cancer chemotherapy on cognitive function from the older patient's perspective. *Breast Cancer Research and Treatment*, *98*, 343–348.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lezak, M., Howieson, D., Loring, D., & Hannay, H.F.J. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Maloni, J.A., Park, S., Anthony, M.K., & Musil, C.M. (2005). Measurement of antepartum depressive symptoms during high-risk pregnancy. *Research in Nursing & Health*, *28*, 16–26.

McGraw, K. & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Parks, K.A., Pardi, A.M., & Bradizza, C.M. (2006). Collecting data on alcohol use and alcohol-related victimization: A comparison of telephone and Web-based survey methods. *Journal of Studies on Alcohol*, *67*, 318–323.

Plassman, B.L., Newman, T.T., Welsh, K.A., Helms, M., & Breitner, J.C.S. (1994). Properties of the telephone interview for cognitive status: Application in epidemiological and longitudinal studies. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *7*, 235–241.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measures*, *1*, 385–401.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie*, *28*, 286–340.

Ross, T.P. (2003). The reliability of cluster and switch scores for the Controlled Oral Word Association Test. *Archives of Clinical Neuropsychology*, *18*, 153–164.

Smith, A. (1982). *Symbol Digit Modalities Test Manual*. Los Angeles, CA: Western Psychological Services.

Squire, L.R. & Zouzounis, J.A. (1988). Self-ratings of memory dysfunction: Different findings in depression and amnesia. *Journal of Clinical & Experimental Neuropsychology*, *10*, 727–738.

Stewart, A., Bielajew, C., Collins, B., Parkinson, M., & Tomiak, E. (2006). A meta-analysis of the neuropsychological effects of adjuvant chemotherapy treatment in women treated for breast cancer. *Clinical Neuropsychologist*, *20*, 76–89.

Tchen, N., Juffs, H.G., Downie, F.P., Yi, Q.L., Hu, H., Chemerynsky, I., Clemons, M., Crump, M., Goss, P.E., Warr, D., Tweedale, M.E., & Tannock, I.F. (2003). Cognitive function, fatigue, and menopausal symptoms in women receiving adjuvant chemotherapy for breast cancer. *Journal of Clinical Oncology*, *21*, 4175–4183.

Tombaugh, T.N. & McIntyre, N.J. (1992). The Mini-Mental State Examination: A comprehensive review. *Journal of the American Geriatrics Society*, *40*, 922–935.

Tulsky, D., Zhu, J., & Ledbetter, M. (1997). *WAIS-III and WMS-III Technical Manual*. San Antonio: The Psychological Corporation.

Uchiyama, C.L., D'Elia, L.F., Dellinger, A.M., Becker, J.T., Selnes, O.A., Wesch, J.E., Chen, B.B., Satz, P., van Gorp, W., & Miller, E.N. (1995). Alternate forms of the Auditory-Verbal Learning Test: Issues of test comparability, longitudinal reliability, and moderating demographic variables. *Archives of Clinical Neuropsychology*, *10*, 133–145.

van Dam, F.S.A.M., Schagen, S.B., Muller, M.J., Boogerd, W., v.d. Wall, E., Droogleever Fortuyn, M.E., & Rodenhuis, S. (1998). Impairment of cognitive function in women receiving adjuvant treatment for high-risk breast cancer: High-dose versus standard-dose chemotherapy. *Journal of the National Cancer Institute*, *90*, 210–218.

Wechsler, D. (1997). *WAIS-III Administration and Scoring Manual*. San Antonio: The Psychological Corporation.

Wefel, J.S., Lenzi, R., Theriault, R.L., Davis, R.N., & Meyers, C.A. (2004). The cognitive sequelae of standard-dose adjuvant chemotherapy in women with breast carcinoma: Results of a prospective, randomized, longitudinal trial. *Cancer*, *100*, 2292–2299.

Welsh, K.A., Breitner, J.C.S., & Magruder-Habib, K.M. (1993). Detection of dementia in the elderly using the Telephone Screening of Cognitive Status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *6*, 103–110.

Zakzanis, K., Leach, L., & Kaplan, E. (1999). *Neuropsychological differential diagnosis*. Lisse, Netherlands: Swets & Zeitlinger Publishers.