MEDICAL CARE
Volume 41, Number 7, pp 853–858

Test-retest Reliability of a Complex Human Immunodeficiency Virus Research Questionnaire Administered by an Audio Computer-assisted Self-interviewing System

Christopher S. Krawczyk, MPH,* Lytt I. Gardner, PhD,* Jichuan Wang, PhD,† Ramses Sadek, PhD,* Anita M. Loughlin, PhD,‡ Pamela Anderson-Mahoney, PhD,§ Lisa Metsch, PhD,‖ and Sonya Green, BA,¶
for the Antiretroviral Treatment and Access Study Group

Objectives. To evaluate the test-retest reliability of a complex questionnaire administered by Audio Computer-assisted Self-interviewing to recently diagnosed human immunodeficiency virus-positive patients.

Methods. Thirty-seven English-speaking and 32 Spanish-speaking participants completed both test and retest interviews. Pearson correlation coefficients (r) and kappa (κ) and weighted kappa (κ) statistics were obtained for individual questions. From these, overall κ and Pearson correlation coefficients were calculated across all variables and for groups of questions.

Results. Overall measures of reliability were $\kappa = 0.767$, $r = 0.728$. Some variation in reliability existed for different response formats, question content groups, and languages of the participants. Differences in overall reliability by Spanish compared with English participants were small and not statistically significant.

Conclusions. Audio Computer-assisted Self-interviewing provides reliable measures for items assessed in the Antiretroviral Treatment and Access Study baseline questionnaire. Some differences exist as a result of question content, interview language, and response format, requiring assessment in future studies and consideration in designing Audio Computer-assisted Self-interviewing systems and questionnaires.

Key words: HIV; ACASI; reliability; survey; design. (Med Care 2003;41:853–858)

Human immunodeficiency virus (HIV) research frequently requires the collection of sensitive, personal information from study participants such as sexual risk behaviors and involvement with the use, sale, or sharing of illegal drugs or drug works. Thus, HIV study participants are often concerned with confidentiality, may be reluctant to disclose this information to researchers, or may be inclined to provide socially acceptable responses. As a result, there

From the *Centers for Disease Control and Prevention, Division of HIV/AIDS Prevention, Atlanta, Georgia.

From †Wright State University, School of Medicine, Center for Interventions, Treatment and Addictions Research Department of Community Health, Dayton, Ohio.

From the ‡Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland.

From the §Health Research Association, Los Angeles, California.

From the ‖University of Miami School of Medicine, Department of Epidemiology and Public Health, Miami, Florida.

From the ¶Emory University School of Medicine, Department of Medicine, Division of Infectious Diseases, Atlanta, Georgia.

Supported in part by a cooperative agreement from the Centers for Disease Control and Prevention.

Address correspondence and reprint requests to: Lytt I. Gardner, PhD, Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Mailstop E-45, Atlanta, GA 30333. E-mail: lig0@cdc.gov

is an increased potential for the underreporting of various behaviors.

Recently, a number of studies have analyzed different interviewing techniques aimed at reducing underreporting and providing a more confidential interviewing environment. These techniques include self-administered pencil and paper interviews,[1,2] audiocassette-assisted self-administered questionnaires,[3] and Audio Computer-assisted Self-interviewing (ACASI). Like self-administered interviews, ACASI can offer a confidential environment for respondents and allows them to answer at their own pace. Additionally, ACASI reduces literacy concerns, automates complex skip patterns, provides a uniform system for collecting data, and eliminates the need for paid data entry and verification.

Studies assessing the utility of ACASI have indicated increases in event reporting,[4–7] improved validity,[8] and improved reliability compared with personal interviews[9] and self-administered interviews.[10] However, evidence suggests impairments of reliability related to questionnaire design,[9] and no studies have assessed the effects of administering interviews in multiple languages. The purpose of the present study is to evaluate the reliability of a complex ACASI-administered questionnaire under the hypothesis that no difference in reliability exists by response format, question content, or language of administration.

Methods

The Antiretroviral Treatment and Access Study (ARTAS) funded by the Centers for Disease Control is a randomized controlled trial evaluating the benefits of an intensive case management intervention on improving access and linkage to HIV medical care.[11,12] This study is being conducted in four US metropolitan locations: Atlanta, Baltimore, Los Angeles, and Miami.

Before the case management intervention, HIV-positive participants were recruited from all four ARTAS study sites for the current study. Two study sites (Atlanta and Baltimore) recruited English-speaking participants, whereas the other two study sites (Los Angeles and Miami) recruited Spanish-speaking participants. Local and Centers for Disease Control institutional review boards reviewed and approved all ARTAS protocols and questionnaires.

English or Spanish versions of the ARTAS baseline questionnaire were administered via ACASI (Questionnaire Development System, version 1.19, NOVA Research Co., Bethesda, MD). Questions were in sections of demographics, medical history, HIV medical care, the participant's HIV disease (HIV-related morbidity, knowledge, and attitude), sexual risk behaviors, substance use, and perceived social support.

Test-retest reliability in this study was the comparison of results from test variables with retest variables given at two separate interviews conducted 72 hours apart. This timeframe is long enough to reduce the effects of recalling test answers and short enough to reduce, although not entirely eliminate, the likelihood of true changes occurring between test and retest.

Pearson correlation coefficients on continuous variables were used to determine whether test and retest measures have identical true scores and equal error variances.[13–15] Reliability of categorical (nominal or ordinal) variables was measured using κ statistics. Zero cells were replaced with a weight of 1E-8. Individual κ values for independent variables were then combined into overall κ values[15] for all variables, the question content categories, and response formats (Table 1) calculated as follows:

$$\kappa_{\text{overall}} = \frac{\sum_{m=1}^{g} \kappa_m / V_m(\kappa_m)}{\sum_{m=1}^{g} 1 / V_m(\kappa_m)}$$

where g is the total number of questions in a group and $V_m(\kappa_m)$ is the variance of κ for question m.

When calculating overall κ values for a given group of questions, all estimated κ values with a value of 1.0 (and zero variance) were excluded. This step produced conservative overall κ estimates and prevented inflated overall κ estimates.

Overall Pearson correlation coefficients were weighted averages where the weight is the number of participants responding to a variable divided by the total number of participant responses in a content category, response format category, or all variables.

To assess the effect interview language had on reliability, overall κ and Pearson correlation coefficients by language of interview were obtained.

TABLE 1. Examples from the Antiretroviral Treatment and Access Study Instrument: Question Content Groups and Response Formats

Question Content Category	Example Question	Example Responses	Response Format Category
Demographics	What was your individual income last year before taxes?	1 = less than $5000 2 = $5001 to $10,000 3 = $10,001 to $20,000 4 = $20,001 to $30,000 5 = $30,001 to $50,000 6 = $50,001 or more	Multiple response
Social support/legal	Do you attend any support groups for HIV/AIDS?	1 = Yes 2 = No	Dichotomous
HIV risk behaviors	How many days in the last 30 days have you injected heroin?	(Please give the number of days)	Continuous
Medical care	When was your most recent T-cell (CD4) count?	— —/— — — — m m y y y y	Date

Overall κ and Pearson correlation coefficients were also calculated within subgroups of response format and question content.

Results

Sixty-nine participants completed both test and retest. Of these 69 participants, 37 (53.6%) completed the English interview, and 32 (46.4%) completed the Spanish interview. Participants ranged in age from 18 to 64 years, with the mean age 40.9 years (SD, 9.6). Demographically, participants were primarily single (50.7%), male (65.2%), and black and non-Hispanic (50.7%) or Hispanic (43.5%); 56 (81.2%) had reported annual incomes below $10,000, and 43 (62.3%) had less than a high school graduate education. English-speaking and Spanish-speaking participants were demographically similar for age ($P = 0.58$), marital status ($P = 0.57$), gender ($P = 0.57$), income ($P = 0.76$), and education ($P = 0.31$).

A total of 205 applicable variables were tested and retested. Statistics for κ could not be obtained for 13 (7.5%) of 174 categorical variables because of small numbers. A further six of the 174 variables with κ values of 1.0 were excluded. Overall reliability measures were 0.767 for κ and 0.728 for Pearson correlation coefficients (Table 2).

Analyses of response format and question content indicated only slight differences between multiple response and dichotomous response formats, whereas larger differences existed between continuous responses and date variables. Reliability for the content groups was less consistent as demographic and HIV risk behavior questions were more reliable than medical care, health and social support, or legal questions. Stratified analyses indicated little consistency regarding the effects of response format within content groups (eg, greater use of dichotomous responses did not yield greater overall reliability regardless of content group).

The overall reliability of English-administered and Spanish-administered questionnaires was similar regardless of measurement (Table 3). Analyses of response format and question content produced mixed results. However, these results suffered from the effects of shrinking numbers (both intravariable and intrasubgroup) because of further stratification.

Discussion

This study indicates ACASI provides reliable measures for items assessed in the ARTAS baseline questionnaire, with some differences caused by question content and response format and little difference caused by interview language.

One of the key issues addressed by the current study is the effect of response format. Although researchers should limit the complexity of interview questions, some items require more complex reporting measures to obtain the most useful and complete data. In our questionnaire, a large number of questions required respondents to choose from multiple possible responses; to recall fre-

TABLE 2. Measures of Reliability: Overall κ and Weighted Averages for Pearson Correlation Coefficients

	Overall κ		Pearson Correlation Coefficient	
Categroy	n	Value	n	Value
All variables	155	0.767	31	0.728
Continuous			21	0.781
Dates			10	0.622
Multiple response:	76	0.752		
>4 Categories	55	0.748		
≤4 Categories	21	0.764		
Dichotomous	79	0.785		
HIV risk behaviors by	18	0.839	16	0.763
Continuous			16	0.763
Multiple response	2	0.951		
Dichotomous	16	0.803		
Medical care/health by	53	0.673	8	0.528
Dates			8	0.528
Multiple response	42	0.685		
Dichotomous	11	0.578		
Social support/legal by	65	0.693	2	0.738
Continuous			2	0.738
Multiple response	21	0.632		
Dichotomous	44	0.716		
Demographics by	19	0.890	5	0.940
Continuous			3	0.899
Dates			2	0.999
Multiple response	11	0.868		
Dichotomous	8	0.919		

quencies of events in the past 30 days, 6 months, or 1 year; or to report the month and year of a specific event. Among these, only questions asking for dates exhibited below-adequate reliability. Although the reasons for reduced reliability of date measurements need further exploration, the overall findings are encouraging considering the response characteristics of the current questionnaire, and should encourage further use of ACASI even in complex research settings.

Closer inspection of the date data revealed large recall errors (test-retest mean difference in days for six nonreliable date variables; mean [SD]: 44.4 [69.4], 52.0 [83.7], 53.3 [96.5], 54.8 [135.1], 197.2 [566.1], 730.5 [1008.1]), which was expected, because small differences in individual answers would not dramatically lower the overall correlation. Although the confidential environment created by ACASI benefits much of data collection, the quality of date data may suffer from this environment mainly because ACASI systems are currently unable to replicate the probing abilities of research interviewers. Development of ACASI systems with visual aids such as drop-down calendars may help improve recall and reliability. Researchers may also consider asking date questions in a different manner, such as, "Did you test positive for HIV in 2000? In 2001?" instead of, "When did you test positive for HIV?"

Results from this study also indicate that language of interview had little effect on reliability. These results are critical because many studies, not just those in HIV-infected persons, involve multilingual populations. We expected any differences caused by language to occur primarily with the Spanish instrument because of problems with the translation from English. Although some errors caused by translation are possible, these appear to have had little effect on overall reliability.

Two limitations of these analyses are the small number of responses to some variables and the small number of variables used in some overall calculations. In stratified analyses, some variables had minimal responses resulting in undefined measures of κ. Additionally, as the number of variables used in overall measurements decreased, the variance of those estimates increased, resulting in less precise

TABLE 3 Measures of Reliability, English-speaking and Spanish-speaking Participants: Overall κ and Weighted Averages for Pearson Correlation Coefficients

	Overall κ				Pearson Correlation Coefficient			
	English Participants		Spanish Participants		English Participants		Spanish Participants	
Category	n	Value	n	Value	n	Value	n	Value
All variables	131	0.729	126	0.714	31	0.6960	31	0.7176
Continuous					21	0.7045	21	0.8061
Dates					10	0.6791	10	0.5407
Multiple response	66	0.734	57	0.716				
>4 Categories	48	0.748	43	0.683				
≤4 Categories	18	0.673	14	0.788				
Dichotomous	65	0.720	69	0.712				
HIV risk behaviors	17	0.784	15	0.775	16	0.6964	16	0.7895
Medical Care/health	46	0.699	40	0.638	8	0.5989	8	0.4260
Social support/legal	53	0.642	54	0.633	2	0.4253	2	0.8450
Demographics	15	0.859	17	0.837	5	0.9586	5	0.9178

measures of κ. Therefore, stratified analyses of interview language and response format and content group should be considered skeptically, and more emphasis should be placed on analyses of interview language and all variables combined. Another concern is the effect of high-expected agreement and estimating κ, particularly for dichotomous items. Our estimates of κ do not exclude variables with high expected agreement. Therefore, our results likely underestimate the true value of κ.

No mode of data collection is obstacle-free, including ACASI. However, a number of recent studies have shown that ACASI improves respondent confidentiality and feelings of privacy, provides a standard mode of data collection, addresses literacy concerns, and decreases underreporting of events. Additional research, including the present study, has concluded that ACASI provides reliable measures, particularly when the effects of questionnaire design are properly addressed. Our results using ACASI should encourage its increased use in HIV studies and other studies with personally sensitive questions, particularly as this technology further evolves and improves.

Acknowledgments

The ARTAS study group consists of the authors and Dr. Lytt I. Gardner, Dr. Scott Holmberg, and Christopher S. Krawczyk from the Centers for Disease Control and Prevention; Dr. Carlos del Rio, Sonya Green, Maribel Barragan, Dr. Wayne Duffus, Dr. Michael Leonard, Catherine Abrams, Christine O'Daniels, Felicia Berry, Valarie Hunter, and Howard Pope from the Emory University School of Medicine, Division of Infectious Diseases; Dr. Steffanie Strathdee, Anita Loughlin, Steven Huettner, Matt Woolf, Ovedia Burt, and Janet Reaves from the Johns Hopkins Bloomberg School of Public Health; Dr. Peter Kerndt, Dr. Pamela Anderson-Mahoney, Bobby Gatson, Brandon Schmidt, Norma Perez, Amy Chan, Lawrence Fernandez, Jr., Shannon Curreri, Eric Valera, Stella Gutierrez, Dr. Sharon Hudson, and Tony Wafford from Health Research Association; Dr. Lisa Metsch, Dr. Clyde McCoy, Dr. Gordon Dickinson, Dr. Toye Brewer, Dr. Orlando Gomez, Eduardo Valverde, Brad Wohler-Torres, Lauren Gooden, Dr. Wei Zhao, Faye Yeomans, and Yolanda Davis Camacho from the University of Miami School of Medicine; and Dr. Harvey A. Siegel, Richard C. Rapp, Teri L. Rust, and Dr. Jichuan Wang from the Wright State University School of Medicine.

References

1. **Aquilino W.** Interview mode effects in surveys of drug and alcohol use. Public Opinion Q 1994;58:210–240.

2. **Aquilino W, LoSciuto L.** Effects of interview mode on self-reported drug use. Public Opinion Q 1990;54:362–395.

3. **Boekeloo BO, Schiavo L, Rabin DL, et al.** Self-reports of HIV risk factors by patients at a sexually transmitted disease clinic: audio vs. written questionnaires. Am J Public Health 1994;84:754–760.

4. **Des Jarlais DC, Paone D, Milliken J, et al.** Audio-computer interviewing to measure risk behaviour

for HIV among injecting drug users: a quasi-randomised trial. Lancet 1999;353:1657–1661.

5. **Metzger DS, Koblin B, Turner C, et al.** Randomized controlled trial of audio computer-assisted self-interviewing: utility and acceptability in longitudinal studies. Am J Epidemiol 2000;152:99–106.

6. **Turner CF, Ku L, Rogers SM, et al.** Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. Science 1998;280:867–873.

7. **Locke SE, Kowaloff HB, Hoff RG, et al.** Computer-based interview for screening blood donors for risk of HIV transmission. JAMA 1992;268:1301–1305.

8. **Murphy DA, Durako S, Muenz LR, et al.** Marijuana use among HIV-positive and high-risk adolescents: a comparison of self-report through audio computer-assisted self-administered interviewing and urinalysis. Am J Epidemiol 2000;152:805–813.

9. **Williams ML, Freeman RC, Bowen AM, et al.** A comparison of the reliability of self-reported drug use and sexual behaviors using computer-assisted versus face-to-face interviewing. AIDS Educ Prev 2000;12:199–213.

10. **Navaline HA, Snider EC, Petro CJ, et al.** An automated version of the risk assessment battery (RAB): enhancing the assessment of risk behaviors. AIDS Res Human Retro 1994;10(Suppl 2):S281–S283.

11. **Anderson-Mahoney P, Green S, Loughlin AM, et al.** Assessments and interventions to improve access and use of HIV care in newly diagnosed HIV-infected persons: Antiretroviral Treatment and Access Studies (ARTAS). In: 2001 National HIV Prevention Conference Program and Abstract Book. Atlanta, GA: Centers for Disease Control and Prevention. 2001:493.

12. **Rapp RC, Siegal HA, Valverde E, et al.** Report on the pilot phase of the Antiretroviral Treatment and Access Studies (ARTAS): an emphasis on ARTAS case management. In: 2001 National HIV Prevention Conference Program and Abstract Book. Atlanta, GA: Centers for Disease Control and Prevention. 2001:492.

13. **Carmines EG, Zeller RA.** Reliability and Validity Assessment. Newbury Park, CA: Sage Publications; 1979.

14. **Bohrnstedt GW.** Measurement. In: Handbook of Survey Research. New York: Academic Press; 1983.

15. **Fleiss JL.** Statistical Methods for Rates and Proportions. 2nd ed. New York: John Wiley & Sons; 1981.