Surface. E. A., Harman, R. P., Watson, A. M., & Thompson, L. F. (2009, April). *Are human- and computer-administered interviews comparable?* Paper to be presented at the 24th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Are Human- and Computer-Administered Interviews Comparable?

Eric A. Surface
SWA Consulting Inc.

Reanna Poncheri Harman
SWA Consulting Inc.

Aaron M. Watson
North Carolina State University/SWA Consulting Inc.

Lori Foster Thompson
North Carolina State University

This field study examined the comparability of interviews administered by humans versus computers (i.e., embodied agents). Ninety-six Korean personnel completed both a human- and a computer-administered interview, counterbalanced to control for order effects and scored by multiple certified raters. Results indicated that the two interview formats exhibited comparable measurement properties.

According to a 2001 survey conducted by American Management Association (AMA), 68% of employers use some form of job skill testing. As such, employment testing is a central part of organizational life and therefore a key area of interest for industrial/organizational (I/O) psychologists. Although most often encountered in the context of employee selection, the term employment testing applies broadly to many techniques used in organizational decision-making. The *Uniform Guidelines on Employee Selection Procedures* (1978) indicates that employment decisions include those associated with hiring, promotion, demotion, membership, referral, retention, and licensing/certification. Given the prevalence of employment testing in the workplace, the need for cost-effective, reliable, and valid assessments is clear.

In the current study, we examine testing in the context of employee certification. Specifically, our focus is on oral proficiency interviewing which is used widely by business, government, and non-profit organizations to assess foreign language proficiency (Swender, 2003). The goal of this study is to examine the comparability of human- and computer-administered versions of this assessment in an effort to determine if the computer-administered version can be used as a reliable and valid replacement for the human-administered version. Evidence of comparability between these two interview modalities has broad implications for the advancement of employment testing.

Oral Examinations and Interviews

Oral exams, or interviews, are an integral part of personnel assessment. For example, the walk-through performance test is an exam in which interviewees are asked to describe in detail the step-by-step process through which job-related tasks are completed (Hedge &

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

Teachout, 1992). Hedge and Teachout found the results of this type of interview to be comparable with that of hands-on work sample tests.

A more common example is the employment interview, which can range from unstructured (i.e., unstandardized interview questions and response scoring) to structured (i.e., standardized interview questions and response scoring; see Huffcutt & Arthur, 1994), with semi-structured interviews (Kohn and Dipboye, 1998) falling between the two extremes. Past research suggests increasing interview structure can enhance the measurement properties (e.g., Campion, Pursell, & Brown, 1988) and criterion-related validity (e.g., McDaniel, Whetzel, Schmidt, & Maurer, 1994) of interviews.

A third example is oral proficiency interviewing for foreign language proficiency certification, which can be used for hiring, promotion, compensation, and other administrative purposes (Swender, 2003). Globalization has led to a heightened demand for workers with foreign language proficiency in many sectors of the economy (Rovira, 2003; Weber, 2004). Globalization has also increased the demand for individuals from other countries, such as Korea, to learn English (Faiola, 2004). As a result, there is a demand for assessment of foreign language proficiency in many contexts, including government, military, education and business organizations (Swender, 2003). Speaking proficiency is often assessed using an interview format, in which the interviewer uses prompts to elicit a speech sample from the examinee, which is subsequently rated for proficiency (Breiner-Sanders, Lowe, Miles, & Swender, 1999). These rater-based language proficiency assessments have been found to be psychometrically sound (e.g., Surface & Dierdorff, 2003), and the American Council on Education (ACE) awards college credits on the basis of oral proficiency interview ratings.

Although they provide in-depth information which can aid in personnel assessment, selection, placement, and promotion, using oral exams to assess language skills or any other attributes is an expensive and time consuming proposition. Unlike paper-and-pencil tests that can be mass administered, interviews and other oral exams are usually administered to one interviewee at a time. Often, they require examinees and interviewers to meet at a common physical location in which travel costs such as airfare and lodging are incurred. Additional expense and time are devoted to training interviewers, which is particularly important when highly structured interviews are used. Although research has shown that increased structure enhances interview reliability and validity, there are still many potential sources of error, such as interpersonal biases (e.g., high self-monitors may receive more favorable evaluations; Lazar, Kravetz, & Zinger, 2004; Osborn, Field, & Veres, 1998), when using interviews for assessment. There is no question that interviews will continue to be an integral part of personnel assessment (Kennedy, 1994); however, there is a need to increase the efficiency and cost-effectiveness of interviews.

Technology Driven Interviews

Technology is changing the nature of selection and assessment. With the advent of advanced computing technologies and the Internet, recent years have seen dramatic changes in the way tests and other assessments are administered (Thompson, Surface, Martin, & Sanders, 2003). For example, paper-and-pencil tests have moved to online assessments – some on-site and proctored, others completed remotely from a location of the examinee's choosing (Templer & Lange, 2008).

Technology has also changed the nature of interviews. To circumvent travel expenses and to allow access to a more diverse pool of applicants, several remote interviewing techniques have become available. The telephone (Chapman, Uggerslev, & Webster, 2003; Silvester & Anderson, 2003; Straus, Miles, & Levesque, 2001), videotaped interviews (Van Iddekinge, Raymark, Roth, & Payne, 2006), videoconferences (Chapman & Rowe, 2001, 2002; Chapman et al., 2003; Straus et al., 2001), and instant messaging (Stieger & Göritz, 2006) provide options for remote interviewing. Even immersive multi-user virtual environments such as Second Life have been used as a medium to conduct interviews from a distance (Athavaley, 2007).

Whereas these options reduce travel costs, they still necessitate the time and expense

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

associated with trained interviewers and the logistical constraints associated with scheduling the interviewer and interviewee. Recently, intelligent agent and related technologies (e.g., computer-adaptive testing; Tonidandel, Quiñones, & Adams, 2002) have allowed the removal of human interviewers altogether. Instead of a person behind the avatar, a computer program, known as an embodied agent, asks the interview questions. Computerized versions of the foreign language assessment described above provide a good example of this form of interviewing. Computerized oral proficiency interviews use an Internet-based embodied agent to elicit and collect a ratable sample of speech, eliminating the need for a live interviewer. The speech sample is digitally saved and evaluated later by certified raters, allowing the sample to be scored by certified raters located anywhere in the world (Thompson, Surface, & Whelan, 2007).

With computerized forms of interviewing, options traditionally determined by a human interviewer (e.g., content, ordering, and difficulty of interview questions, presence of follow-up or probing questions) are controlled by a computer program. This interview format has appeal for a number of reasons. It not only reduces travel costs, but it also eliminates the time and expense of paying a trained interviewer as well as the logistical constraints of scheduling an interview. In addition, it potentially removes some of the interpersonal biases that can creep into interviews.

Comparability of Human- and Computer-Administered Interviews

Despite its presumed advantages, the introduction of this new interview type raises questions about its comparability to more traditional interview formats. A fundamental difference between the two formats is the lack of interpersonal interaction that typically takes place during an interview. This generates concerns about (a) interviewee reactions (e.g., perceptions of face validity) to computer-administered interviews as well as (b) the potential impact on measurement and interview performance. While research (e.g., Thompson et al., 2007) has investigated the first of these two issues, no past work has addressed whether removing the human interviewer from the loop affects the construct measurement of an interview. On the one hand, examinees may feel uncomfortable and therefore act unnaturally if the interviewer is not a social actor. On the other hand, theory suggests that people effortlessly and even "mindlessly" apply social expectations and rules to computers (Nass & Moon, 2000). In fact, research has shown that individuals adapt relatively quickly to working with an intelligent agent/robot as a teammate, particularly if the intelligent agent has human-like characteristics (Hinds, Roberts, & Jones, 2004; Nass, Fogg, & Moon, 1996), thereby implying that human- and computer-administered interviews should be comparable.

At present, this issue of measurement equivalence awaits empirical investigation. The purpose of the present study is to provide a rigorous examination of the following questions:

> RQ1: Do multiple raters demonstrate similar conceptualizations (or shared mental models) of an underlying construct assessed by human- versus computer-administered interviews?
>
> RQ2: Are human- and computer-administered interviews rated by multiple raters characterized by comparable levels of reliability?
>
> RQ3: Do human- and computer-administered interviews rated by multiple raters exhibit comparable rating outcomes?

Method

Participants

A sample of 100 participants was randomly selected for this study from the workforce of a company in Korea; 99 of these individuals participated. Thirty-seven percent of the final sample was male and 69% indicated their highest level of education completed was a B.A./B.S. degree. The majority of participants indicated they first studied English in middle school (57%) or primary school (39%). Due to technical issues, speech samples from three computer-administered interviews were un-

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

ratable, reducing the final sample size to 96 participants.

Field Study Design

This field study used a within-subjects design in which each participant completed both a human- and a computer-administered interview. As shown in Figure 1, participants completed a pre-assessment survey, their first interview, their second interview, and a post-assessment survey within a specified time frame. The administration order of the human- and computer-interview formats was counterbalanced to control for order effects, with participants randomly assigned to one of the two interview administration orders.

Interview Formats

Human interviewer. The American Council on the Teaching of Foreign Languages' Oral Proficiency Interview (ACTFL OPI®) is a standardized assessment of speaking proficiency. The assessment is administered as a face-to-face or telephonic interview in which a certified tester—serving as the interviewer—assesses an examinee's speaking proficiency by asking a series of questions in the context of a structured conversation. The question content is based on the examinee's interests as determined by a preliminary set of questions in the interview and is adapted during the interview based on the individual's proficiency level. Each ACTFL OPI® is conducted and rated by certified testers. The interviews are recorded and typically rated by two certified testers—one who interviews the individual and rates the sample after the interview and one who serves as a rater only. The ACTFL testers compare the interviewee's responses with criteria for ten proficiency levels ranging from *Novice Low* to *Superior*, specified in the *ACTFL Proficiency Guidelines – Speaking: Revised 1999* (Breiner-Sanders et al., 2000). Previous research has produced support for the validity (Dandonoli & Henning, 1990) and reliability (Surface & Dierdorff, 2003; Thompson, 1995) of the ACTFL OPI® construct.

Computer interviewer. The ACTFL OPIc® (OPI computer) is a test of spoken English proficiency designed to elicit a sample of speech via computer-delivered prompts (delivered by an embodied agent named Ava). An individual is able to access this ACTFL OPI®-like test without the presence of a live tester to conduct the interview. The range of proficiency assessed by this test is *Novice Low* to *Advanced.* The ACTFL OPIc® uses the same guidelines, protocols, and scale as the ACTFL OPI®. Each test is individualized. An algorithm selects prompts (i.e., interview questions) at random from a database of thousands of prompts. The task and topic areas of these prompts correspond to the test taker's self-reported (via survey) linguistic, interest, and experience profiles. The approximate test time is 20-30 minutes, depending on the level of proficiency of the test taker. The speech sample is digitally saved and rated by certified ACTFL OPIc® raters.

Raters and Rating Protocol

The validity and reliability of a rater-based assessment are a function of the raters applying a shared mental model consistently. Nine raters were recruited from a pool of experienced individuals (certified for rating the human-interview format) and were trained and certified to rate the samples produced by the computer-administered interview. Five of the nine raters were randomly assigned to evaluate each interviewee's speech samples. For a given interviewee, one of the five was assigned to rate both the human and the computerized interviews and the remaining four rated the speech sample from only one of the participant's two interviews. Thus, three individuals rated each participant's human interview and three rated each participant's computerized interview, resulting in 6 ratings and 5 unique raters per interviewee. All raters adhered to the content rating protocols and guidelines normally followed for ACTFL speaking proficiency assessments. As mentioned, each speech sample was evaluated by three raters. Although the typical protocol is to use two raters with a third called upon to arbitrate disagreements, employing three raters allowed for more sophisticated statistical analyses, such as confirmatory factor analysis (CFA).

Analytic Strategy

CFA was used to assess whether or not raters held similar conceptualizations of the underlying construct across computer- and human-administered interviews and to determine the relationship between the rating outcomes of both modalities. Guidance in the measurement invariance/equivalence (MI/E) literature (e.g., Vandenberg & Lance, 2000) was followed. To determine whether or not raters exhibited comparable levels of reliability across interview formats, interrater reliability was calculated using intraclass correlations (ICC; Shrout & Fleiss, 1979), and interrater agreement was calculated using r_{wg} (James, Demaree, & Wolf, 1984).

Results

Interrater Consistency and Agreement

To assess interrater consistency, ICCs were calculated across the three rater positions for both interview modalities and are presented in Table 1. Indices of consistency at or above .70 are traditionally considered to be sufficient (LeBreton, Burgess, Kaiser, Atchley, & James 2003). ICCs for both interview modalities exceeded the minimum standard of .70 and were nearly identical in magnitude.

Interrater agreement was assessed using the r_{wg} index for a single-item measure (see James et al., 1984), which compares observed variance in ratings across multiple raters to that which would be expected by chance. Agreement indices were calculated for each set of ratings provided for each interviewee. Summary statistics for observed r_{wg} values are provided in Table 1. Agreement in both computer- and human-administered interview modalities were comparable and high (median r_{wg} = 1.00 for both modalities), exceeding the critical values for statistical significance (i.e., statistically different than zero) proposed by Dunlap, Burkey, and Smith-Crowe (2003).

Invariance Tests

CFA was employed to assess (a) the equivalence of measurement properties between the two interview modalities, and (b) the relationship between rating outcomes from both modalities. Nested model comparisons were conducted, with each subsequent model imposing additional constraints holding specific measurement properties invariant across the human- and computer-administered interview events (Taris, Bok, & Meijer, 1998; Vandenberg & Lance, 2000). The present analysis adopted a longitudinal measurement invariance approach, assessing the stability of latent constructs across measurement occasions (e.g., Taris et al., 1998).

In the first model, the two speaking proficiency factors were specified as latent variables using their corresponding raters as indicators. In this baseline model, factor loadings, indicator (i.e., observed rating) intercepts, and factor variances were freely estimated across factors. In order to achieve model identification and scaling, factor loadings and indicator intercepts were fixed for the common rater position (i.e., same rater for a given interviewee across interview modalities) for each factor. As is commonly done in longitudinal invariance testing, the error terms associated with the common rater indicators were allowed to correlate across interview modalities (Ployhart & Oswald, 2004; Vandenberg & Lance, 2000). This model represents a test of the equality of factor structure across groups (i.e., configural invariance; Horn & McArdle, 1992). Results indicated adequate model fit for the configural model (see Table 2) and are presented in Figure 2.

To test for full metric invariance, the second model (Model 2) imposed equality constraints on all factor loadings (the Λ_x matrix) across the computer- and human-administered interview occasions. In addition, Model 2 imposed equality constraints across raters within interview modality. This was achieved by fixing the factor loadings of all indicators to unity to equal that of the referent indicator. As differences in chi-square values for large samples approximate the chi-square distribution, chi-square difference ($\Delta\chi^2$) tests were used in model comparisons. Comparison of Model 2 to Model 1 did not produce a significant decrement in fit (see Table 2), demonstrating indicator loadings were invariant across interview modalities.

To test for scalar invariance, Model 3 imposed equality constraints on all indicator intercepts (the τ_x matrix) across the computer- and human-administered interview occasions. In addition, Model 3 imposed equality constraints across raters within interview

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

modality. This was achieved by fixing the intercepts of all indicators to zero to equal that of the referent indicator. Comparison of Model 3 to Model 2 did not produce a significant decrement in fit (see Table 2), demonstrating indicator intercepts were invariant across interview modalities.

Model 4 tested for equality of factor variances between the computer- and human-administered latent variables. If raters' underlying conceptualization of the target construct was unchanged by interview modality, factor variances should be invariant across measurement occasions (Taris et al., 1998). Comparison of Model 4 to Model 3 did not produce a significant decrement in fit (see Table 2), indicating factor variances were invariant across interview modalities.

To test for equality of error variances, Model 5 imposed equality constraints on all indicator uniqueness terms across raters and interview modalities. Comparison of Model 5 to Model 4 did not produce a significant decrement in fit (see Table 2), indicating error variances were invariant across *all* raters regardless of interview modalities. Since factor variances were also constrained to be equal, Model 5 represents a test for invariant indicator reliabilities across interview modalities. Thus, the lack of significant decrement in model fit for Model 5 indicates reliabilities of individual ratings were invariant across *all* raters and interview modalities.

The final model (Model 6) tested for equality of latent factor means across interview modality by imposing equality constraints on both latent means. Results of a model comparison indicated the more parsimonious Model 6 did not significantly degrade model fit. Thus, interviewees' absolute level on the speaking proficiency latent construct was invariant across human- and computer-administered interview modalities at the group level.

Evaluation of Research Questions

RQ1 asked whether multiple raters demonstrate similar conceptualizations (or shared mental models) of the underlying construct across human- and computer-administered interviews. Results indicated ratings within each interview modality loaded onto a common factor, suggesting ratings were indeed indicative of a single underlying construct. Additionally, ratings were found to share a common metric or scale at both the individual rating (metric invariance) and latent factor levels (invariant factor variances). This common metric was found both across raters within interview modality, as well as across modalities. These findings suggest raters possessed a shared mental model of the underlying construct, which did not vary as a function of the interview medium.

RQ2 asked whether human- and computer-administered interviews rated by multiple raters are characterized by comparable levels of reliability. A direct test of the equality of rating reliabilities indicated error associated with individual ratings was invariant across interview modalities (invariant uniquenesses). Also, evidence supporting scalar invariance (Model 3) indicated no detectable systematic rater bias (e.g., leniency, severity, etc.) attributable to the modality of the interview (Vandenberg & Lance, 2000). These findings indicate the reliability of the human- and computer-administered interview assessments were indeed comparable.

RQ3 asked whether human- and computer-administered interviews rated by multiple raters exhibit comparable rating outcomes. Results indicated latent means across the human- and computer-administered interview occasions were invariant. Thus, after accounting for measurement error, no group-level differences in the absolute level of interview ratings were found. Also, the latent factor correlation (.94) was significant and strong, indicating rank-order of interviewees on the latent factor remained largely consistent across interview modality. Thus, rating outcomes appeared highly consistent across human- and computer-administered interviews.

Discussion

As the demand for employment testing increases, organizations will increasingly turn to technology-based assessment solutions. Because employment testing is high-stakes, the use of technology to conduct such testing needs to be thoroughly evaluated. This applies to interviews as well as paper-and-pencil tests. Our field study is an important first step in assessing the measurement comparability of human- and computer-administered interviews. Our findings indicate that raters consistently applied a shared mental model of English speaking proficiency across speech samples from human- and computer-administered interviews. Human- and computer-administered interviews evaluated by multiple raters exhibited comparable levels of reliability, and the interview modality did not appear to affect examinees' performance. Given the importance of and increase in language proficiency assessment, this is good news in that it will improve the efficiency of the testing process (e.g., reducing scheduling complexity) and decrease costs without compromising the measurement properties and effectiveness of the process.

Although our field study included random selection, random assignment, counterbalancing, and other experimental controls, there are a few potential limitations to note—our results may not generalize to other populations, other interview contexts, the measurement of other constructs, or even the measurement of proficiency in languages other than English. Future research should focus on replicating our findings for other populations, interview contexts, constructs, and languages. Despite the potential limitations, our study makes an important contribution to the employment testing literature and serves as a foundation for future research comparing human- and computer-administered interview assessments. To the extent that our finding generalize to other interview-based assessments, organizational practice could change greatly with the adoption of computer-administered interviews, freeing time and resources to focus on other tasks as opposed to scheduling and conducting interviews.

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

References

American Management Association (2001). *2001 AMA survey on workplace testing: Basic skills, job skills, psychological measurement*. Retrieved September 7, 2008 from http://www.amanet.org/research/pdfs/bjp_2001.pdf

Athavaley, A. (2007, June 20). A job interview you don't have to show up for. *Wall Street Journal - Eastern Edition, 249*(143), D1-D8.

Breiner-Sanders, K.E., Lowe, P., Miles, J., & Swender, E. (1999). ACTFL Proficiency Guidelines—Speaking revised 1999. *Foreign Language Annals, 33*, 13–17.

Campion M. A., Purcell E. D., & Brown B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41*, 25-42.

Chapman, D. (2001). The impact of videoconference technology, interview structure, and interviewer gender on interviewer evaluations in the employment interview: A field experiment. *Journal of Occupational & Organizational Psychology, 74*, 279-298.

Chapman, D. S., & Rowe, P. M. (2001). The impact of videoconference technology, interview structure, and interviewer gender on interviewer evaluations in the employment interview: A field experiment. *Journal of Occupational and Organizational Psychology, 74*, 279-298.

Chapman, D. S., & Rowe, P. M. (2002). The influence of videoconference technology and interview structure on the recruiting function of the employment interview: A field experiment. *International Journal of Selection and Assessment, 10*, 185-197.

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology, 88*, 944-953.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL Oral Proficiency Guidelines and Oral Interview Procedure. *Foreign Language Annals, 23*, 11–22.

Dunlap, W. P., Burkey, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for rwg and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*, 356-362.

Equal Employment Opportunity Commission, U. S. Civil Service Commission, U. S. Department of Labor, & U. S. Department of Justice (1978, August 25). Uniform guidelines on employment selection procedures. *Federal Register, 43*, 38290-38309.

Faiola, A. (2004, November 18). English camps reflect S. Korean ambitions; Youth pushed to master 'global language'. *The Washington Post*, p. A.25.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. *Journal of Applied Psychology, 77*, 453-461.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction, 19*, 151-181.

Horn, J., & McArdle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144.

Huffcutt A. I., & Arthur W. Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184-190.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85-98.

Kennedy, R. B. (1994). The employment interview. *Journal of Employment Counseling, 31*, 110-114.

Kohn L. S., & Dipboye R. L. (1998). The effects of interview structure on recruiting outcomes. *Journal of Applied Social Psychology, 28*, 821-843.

Lazar, A., Kravetz, S., & Zinger, A. (2004). Moderating effects of rater personality on the relation between candidate self-monitoring and selection interview ratings. *International Journal of Selection and Assessment, 12*, 321-326.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80-128.

McDaniel M. A., Whetzel D. L., Schmidt F. L., & Maurer S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599-616.

Nass C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*, 669-678.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*, 81-103.

Osborn, S., Field, H., & Veres, J. (1998). Introversion-extraversion, self-monitoring and applicant performance in a situational panel interview: A field study. Journal *of Business and Psychology, 13*, 143-156.

Ployhart, R., & Oswald, F. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, *7*, 27-65.

Rovira, J. D. (2003). Importance of foreign language capabilities. (ALMAR 072/03). Washington, DC: Author. Retrieved September 9, 2005, from http://www.marines.mil/almars/almar2000.nsf/1babcf316f87f38c852569b8008017e7/952d4b8516b446a585256df8006a75f9?OpenDocument

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420- 428.

Silvester, J., & Anderson, N. (2003). Technology and discourse: A comparison of face-to-face and telephone employment interviews. *International Journal of Selection & Assessment*, *11*, 206-214.

Stieger, S., & Göritz, A. (2006). Using instant messaging for internet-based interviews. *CyberPsychology & Behavior, 9*, 552-559.

Straus, S., Miles, J., & Levesque, L. (2001). The effects of videoconference, telephone, and face-to-face media on interviewer and applicant judgments in employment interviews. *Journal of Management*, *27*, 363-381.

Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL oral proficiency interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, *36*, 507-519.

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. *Foreign Language Annals, 36*, 520-526.

Taris, T., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology*, *132*, 301-316.

Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unprocotored field conditions. *Computers in Human Behavior, 24*, 1216-1228.

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, and Spanish. *Foreign Language Annals, 28*, 407-422.

Thompson, L. F., Surface, E. A., Martin, D. L., & Sanders, M. G. (2003). From paper to pixels: Moving personnel surveys to the Web. *Personnel Psychology, 56*, 197-227.

Thompson, L. F., Surface, E. A., & Whelan, T. J. (2007, April). Examinees' reactions to computer-based versus telephonic oral proficiency interviews. Paper presented at the 22nd annual conference of the Society for Industrial and Organizational Psychology, New York, NY.

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

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69.

Van Iddekinge, C. H., Raymark, P. H., Roth, P. L., & Payne, H. S. (2006). Comparing the psychometric characteristics of ratings of face-to-face and videotaped structured interviews. *International Journal of Selection and Assessment, 14*, 347-359.

Weber, G. (2004). English rules. *Workforce Management, 83*, p. 47-51.

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

Table 1. Interrater Consistency and Agreement Indices

	Interrater Consistency			Interrater Agreement		
		95% CI for ICC				
Interview Medium	ICC	Lower	Upper	Median r_{wg}	Min r_{wg}	Max r_{wg}
Human-administered	0.93*	0.90	0.95	1.00*	0.88	1.00
Computer-administered	0.94*	0.92	0.96	1.00*	0.67	1.00

* $p < .05$

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

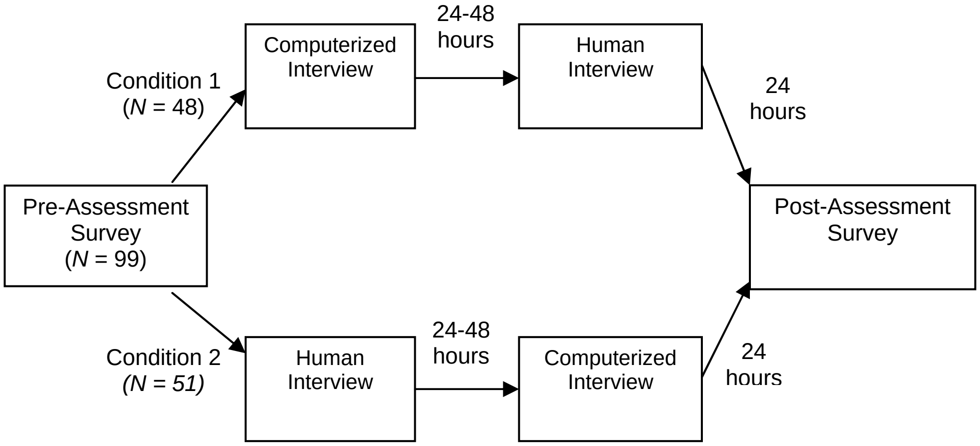
Table 2. Tests of Measurement Invariance Across Computer- and Human-Administered Interviews

Model	χ^2	*df*	Comparison Model	Δχ2	Δ*df*	CFI	TLI	RMSEA	SRMR
1 Configural invariance	4.95	7	-	-	-	1.000	1.000	0.000	0.005
2 Metric invariance ($\Lambda^g = \Lambda^{g'}$)	13.79	11	Model 1	8.84	4	0.997	0.996	0.051	0.037
3 Scalar invariance ($\tau^g = \tau^{g'}$)	15.69	15	Model 2	1.90	4	0.999	0.999	0.022	0.034
4 Invariant factor variances ($\Phi^g = \Phi^{g'}$)	15.92	16	Model 3	0.23	1	1.000	1.000	0.000	0.038
5 Invariant uniquenesses ($\Theta^g = \Theta^{g'}$)	22.71	21	Model 4	6.79	5	0.998	0.999	0.029	0.036
6 Invariant factor means ($\kappa^g = \kappa^{g'}$)	25.75	22	Model 5	3.04	1	0.996	0.998	0.042	0.033

Note. CFI = comparative fit index, TLI = Tucker-Lewis index, RMSEA= root mean squared error of approximation, SRMR= standardized root mean squared residual.

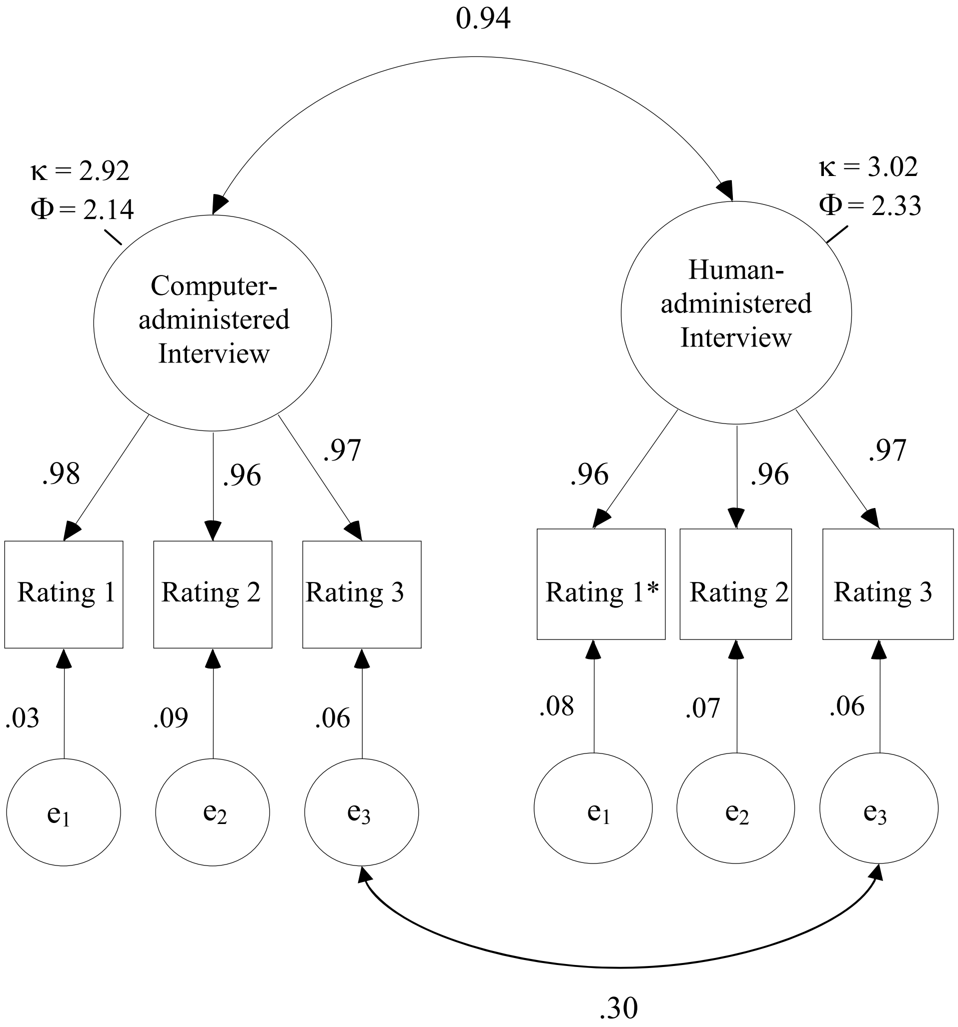
* $p < .05$

Figure 1. Field study design



Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).

Figure 2. Correlated Two-Factor Model (Model 1): Computer- and Human-administered Interview With Correlated Error Terms for Common Rater



Note. Standardized parameter estimates are presented (except for latent means and variances, which are unstandardized). All paths are statistically significant (*p* < .05).
* Indicates the interviewer position for the human-administered interview.

Note. Please do not cite without permission from the authors. Please direct inquiries about this paper to Dr. Eric Surface (esurface@swa-consulting.com).