Psychometric Properties of Automated Video Interview Competency Assessments

Authors:
Joshua Liff[1]
Nathan Mondragon[1]
Cari Gardner[1]
Christopher J. Hartwell[2*]
Adam Bradshaw[1]

[1]HireVue, Inc.; 10876 South River Front Parkway #500, South Jordan, UT 84095
[2]Department of Management, Utah State University; 3555 Old Main Hill, Logan, UT 84322
*Corresponding Author

Author Note:

Portions of this article were originally presented at SIOP, 2022 in Seattle, WA:

Hartwell, C., Liff, J., Gardner, C., & Mondragon, N. (2022). Development and Validation of Asynchronous Competency-based Structured Interview Scoring Algorithms.

We would like to thank Caleb Rottman, Rhett Shipp, and Missy Tracy from HireVue's Data Science team for their close partnership on all of the research to develop and validate the AVI-CAs.

Abstract

Interviews are one of the most widely used selection methods, but their reliability and validity can vary substantially. Further, using human evaluators to rate an interview can be expensive and time-consuming. Interview scoring models have been proposed as a mechanism for reliably, accurately, and efficiently scoring video-based interviews. Yet, there is a lack of clarity and consensus around their psychometric characteristics, primarily driven by a dearth of published empirical research. The goal of this study was to examine the psychometric properties of automated video interview competency assessments (AVI-CAs), which were designed to be highly generalizable (i.e., apply across job roles and organizations). The AVI-CAs developed demonstrated high levels of convergent validity (average *r* value of .66), moderate discriminant relationships (average *r* value of .58), good test-retest reliability (average *r* of .72), and minimal levels of subgroup differences (Cohen's *d*s ≧ -.14). Further, criterion-related validity (uncorrected sample weighted $\bar{r}$ = .24) was demonstrated by applying these AVI-CAs to five organizational samples. Strengths, weaknesses, and future directions for building interview scoring models are also discussed.

Keywords: machine learning, artificial intelligence, video interviews

Psychometric Properties of Automated Video Interview Competency Assessments

Interviews have long been a central part of the hiring process for most organizations. Traditional employment interviews generally consist of one or more interviewers asking questions to, and judging responses from, one or more job applicants. While interviews are typically conducted face-to-face, phone interviews are also used as a less expensive alternative to an in-person interview or as a pre-screen to reduce the number of applicants brought in for face-to-face interviews (Oliphant et al., 2008). Fueled in part by the COVID-19 pandemic when meeting face-to-face was often not advisable (Rubinstein, 2020), video-based interviewing is an interview modality that is rapidly gaining popularity. More recently, organizations and researchers have begun using and studying automated video interviews (AVIs), which are asynchronous video interviews scored by machine learning (ML) algorithms (Hickman et al., 2022). For organizations, there are many potential benefits of using AVIs, including the ability to effectively screen a large number of candidates, reduced administrative costs, less time spent conducting interviews, and the ability to reach a more diverse set of applicants (e.g., Brenner et al., 2016; Chapman & Webster, 2003; HireVue, 2017). While the combination of a technological shift and the COVID-19 pandemic has led to a substantial increase in the use of AVIs in organizations (Lukacik et al., 2022; Tippins et al., 2021) and numerous articles have discussed their potential benefits and drawbacks (e.g., Gonzalez et al., 2019; Tippins et al., 2021), there is a paucity of research that has examined their psychometric properties (Chamorro-Premuzic et al., 2016; Gorman et al., 2018; see Hickman et al., 2022 for an exception). This is problematic as AVIs used in a selection context are subject to both legal regulations and professional testing guidelines such as Title VII of the Civil Rights Act of 1964, the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial Organizational Psychology, 2018), and The Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978).

There are several contributions this study makes to the field of personnel selection. First, it is the first study we are aware of that developed a broad set of AVIs to measure a diverse range of KSAOs across a variety of jobs, organizations, and industries. Having a wide variety of highly generalizable AVIs increases the practical utility of this research as it allows organizations to create an assessment by selecting relevant AVIs based on job characteristics as identified through a job analysis. In addition, as the development of ML models requires large amounts of data, generalizable AVIs are beneficial to organizations which may not have the applicant volumes or the data science expertise to develop job-specific AVIs. Second, it answers calls to examine the psychometric properties of AVIs, which is critical as the adoption of AVIs is outpacing research (Chamorro-Premuzic et al., 2016; Gonzalez et al., 2019; Lukacik et al., 2022; Woods et al., 2019). Third, this study examines the demographic subgroups differences of AVIs. This is important as ethical concerns around the use of AVIs are often raised along with calls for more research (Hunkenschroer & Luetge, 2022; Woods et al., 2019). Finally, while previous research has relied on university students, research panels (e.g., Prolific), or other non-applicant video interview response data to build and validate their AVI models (e.g., Hickman et al., 2022; Köchlig et al., 2020), the present study uses real applicants to build the AVI models and organizational performance metrics to test the AVIs' criterion-related validity. This is an important contribution as the motivation of real applicants, and the resulting behavioral response patterns they exhibit in the video interview, may be different between panel or student samples.

Automated Video Interview (AVI) Background

AVIs are developed by first processing verbal, paraverbal (e.g., pitch, cadence, tone), and nonverbal (e.g., facial expressions) behaviors exhibited in asynchronous video interviews and then using the resulting data as predictors in a ML model to predict the outcome of interest. There are many reasons an organization may implement AVIs, including reduced costs, increased effectiveness, and increased standardization. While potentially beneficial, there are also several drawbacks to using AVIs, including mixed applicant reactions to ML-based decisions (Langer & Landers, 2021) and quickly evolving legislation (e.g., Artificial Intelligence Video Interview Act in the state of Illinois, 2020) which requires organizations to keep apprised of the rapidly changing legal environment.

AVIs are a method (i.e., video interview) which includes a scoring algorithm (i.e., ML model). As such, AVIs can be designed to measure different constructs or KSAOs (Arthur & Villado, 2008). This distinction between method and construct is important as an AVI designed to measure one construct may demonstrate different psychometric properties than an AVI designed to measure a separate construct. Likewise, different decisions in the ML model building process can impact the psychometric properties of AVIs.

Previous research has predominantly focused on the predictive validity of AVIs in respect to predicting big-five personality (Chen et al., 2016; Chen at al., 2018; Hickman et al., 2019; Hickman et al., 2022) and social skills such as communication (Chen et al., 2016; Rasipuram et al., 2016; Torres & Gregory, 2018). This previous research found that it is indeed possible to measure various constructs using AVIs, although the effectiveness is dependent, in part, upon the construct that is being measured. For example, Hickman et al. (2022) found that AVIs designed to measure interviewer ratings of applicant personality were twice as accurate as AVIs designed to measure self-reported personality.

Competency Measurement with AVIs

In the current research, we answer the call for greater examination of the validity of specific interview constructs, versus composite scores, of job interviews (Huffcutt et al., 2001). Accordingly, we decided to create AVIs designed to measure competencies, which are clusters of knowledge, skills, abilities, and other attributes (KSAOs) applied in an organizational or job context that contribute to business success (Campion et. al., 2011; Carroll & McCrackin, 1998; Kessler, 2006; Kochanski, 1997; Sanchez & Levine, 2009). Competency assessment was first introduced in the 1970s (McClelland, 1973) and has been used in a variety of fields, such as medicine, engineering, computer programming, and education (e.g., Arikan et al., 2020; Dano, 2019; Howard, et al., 2018; Ullah et al., 2019).

The principal reason for focusing the content of the AVIs on competencies was because competency frameworks are commonly used within organizations (Green, 1999; Campion et al., 2011), thereby increasing the applicability of this approach in an applied setting. We identified fifteen clusters of KSAOs which we deemed to be most applicable and predictive across entry-level roles through a review of the psychological literature (e.g., Hogan et al., 2013), extant competency frameworks (e.g., Bartram, 2005), O*NET (Peterson et al., 1999), a comprehensive review of job analyses (e.g., use of critical incidents across job analytic studies to inform construction of behaviorally anchored rating scales) and criterion validation studies conducted with the authors' various clients. As such, our objective in the current research is to develop and examine the psychometric properties of fifteen AVI competency assessments (AVI-CAs).

Competencies can be identified and applied across a variety of jobs at different organizational levels as they are tied to business objectives and strategies. For example, Problem Solving may be useful for a customer service representative, a production floor manager, or a company CEO, but the level of Problem Solving needed is likely to change at different levels of the organization. Practically, a competency-based approach is beneficial to organizations in that the competencies measured by our AVI-CAs can often be mapped to pre-existing organization-specific competency frameworks and linked to job analytic results. This results in a highly flexible and scalable assessment approach, ultimately enhancing both job relevance and legal defensibility. Appendix A presents the competency definitions, key behaviors, a sample interview question representative of each competency domain, and theoretical mappings between the AVI competencies and elements in the O*NET Content Model (Peterson et al., 1999) as well as with component competencies (i.e., the finest level of specificity) in The Great Eight Competencies (Bartram, 2005).

To produce the linkages between the competencies in the current study and the O*NET content domain, a content linkage study was conducted with 15 Industrial-Organizational (IO) Psychologists who reviewed AVI competency definitions and O*NET element descriptors to rate the extent to which there was conceptual overlap between the two frameworks. Each O*NET descriptor was compared with the 15 AVI competency definitions and was rated on a 5-point scale ranging from 'Not at all Related' to 'Extremely High Degree of Overlap,' with the latter category indicating when constructs are essentially conceptually equivalent (e.g., the Composure AVI competency area and Stress Tolerance from O*NET's Content Model). Across all 15 O*NET descriptor areas that were examined, very high levels of inter-rater agreement were obtained (see Online Supplement Table S1 for details on interrater agreement across all O*NET descriptor areas). The median Interclass Correlation Coefficient (ICC)[1] was .96 (Range: 0.74 to 0.99) based on the relevance ratings of the 15 O*NET Descriptor Areas compared to all the competencies included in this research study. Thus, evidence suggested IO Psychologists had highly similar judgments of the degree of relevance across all attributes from O*NET that were compared to the AVI competencies. Appendix A presents up to 5 O*NET elements that received a relevance rating of 3 (a moderate degree of overlap) or greater and a standard error of the mean less than or equal to .51 (see Table S1 for the full rating scale used in this linkage study).

To produce the linkages between AVI competencies and the component-level competencies in The Great Eight Competencies framework (Bartram, 2005), two IO Psychologists who were involved in the development of the AVI competency framework independently reviewed both frameworks to compare the conceptual overlap of behaviors specified in the definition of each set of competencies. Similar to the O*NET linkage process, SMEs were able to link multiple Great Eight component-level competencies to an AVI competency if two or more behaviors specified in a definition had conceptual overlap with the AVI definition and/or key behaviors. There was complete agreement between the linkages each SME produced. Overall, AVI competencies are broader than the component-level competencies in the Great Eight. This breadth difference between AVI competencies and The Great Eight Competencies is illustrated in Appendix A since all the AVI competencies have multiple linkages with the Great Eight except for Service Orientation, which was linked to the Focusing on Customer Needs and Satisfaction component-level competency only.

[1] A two-way mixed model ICC based on absolute agreement and an average type model (ICC 2,k) was computed since the average ratings are what are used operationally to link O*NET descriptors with AVI competencies (see Trevethan, 2016, for details on ICC model types).

The KSAOs in the present study occupy a considerable breadth and depth of the construct domain with some of the KSAOs being conceptually similar to personality traits, for example, Dependability (KSAO in current study) is conceptually similar to Conscientiousness (big-five personality). Yet other KSAOs in this research, such as Problem Solving, are conceptually distinct from personality. Figure 1 presents the competency framework in the current study, and it specifies into which of the three different domains each competency conceptually fits: (1) *Work with People* - The extent to which individuals possess the KSAOs to form productive and rewarding relationships with others at work, (2) *Work with Information* - The extent to which individuals have the KSAOs to effectively process information and data they encounter in their work role to drive decisions and action, and (3) *Personality & Work Style* - The extent to which individuals have the appropriate composition and level of personality, motivation, and attitudes to meet the people, data, and information demands of the job specified in other competency domain areas.

Building AVI-CAs

The effectiveness of AVIs is highly dependent upon the underlying data (both the predictors and criterion) along with the natural language processing (NLP) and ML techniques used to build the AVI. While our Study Design and Methods section discusses our data and model building process in detail, we provide a more theoretical discussion of these considerations next to orient the reader and highlight the importance of these research design decisions.

Underlying Data Used to Build AVIs

Similar to traditional assessment or test development methods, one key aspect of AVI model validity is the data and test stimuli (e.g., interview question prompts) used to build the AVI model (Gonzalez et al., 2019). There are four primary characteristics to consider when building an AVI model: interview quality, training sample representativeness, predictor characteristics, and criterion quality. Researchers should consider each of these aspects in the development process, as these data characteristics set the upper limit for the validity of the AVI in different contexts.

Interview quality refers to the questions asked in the video interviews. It is important that the questions asked result in high-quality applicant responses as this generates the predictors that will be used in the ML model. For example, while an AVI may be able to measure extraversion using a wide variety of broad interview questions, the AVI would likely demonstrate higher convergent validity (i.e., the extent to which the AVI is related to other measures of the same construct) when the interview questions are designed to measure extraversion as the question prompts are more likely to elicit stronger trait-relevant responses (Tett & Burnett, 2003). Therefore, the current research used rigorously developed job-relevant past behavioral and situational questions. These structured question types are also situation-specific, either asking for an example of past behavior or a description of the behavior one would take in a hypothetical situation that matches the job context. Thus, these question types elicit demonstration of competencies in an applied setting (Hartwell et al., 2019; Kessler, 2006).

Training sample representativeness refers to the extent to which the data used to train ML algorithms reflects the populations and contexts (e.g., employee selection) to which one intends to generalize results. Since an interview is a performance event, it is important that the sample

used in training reflects the environment in which AVI-CAs will be used. For instance, both the motivation and range of behaviors exhibited is likely to vary between a research panel sample (e.g., Qualtrics) and a sample of actual applicant responses in a selection context. When the training context and application of use vary significantly, there are substantial threats to external validity. Given this, the current research used real applicants applying to a wide variety of jobs and organizations.

One data characteristic is the type of predictors used to build AVIs. In an in-person interview, the interviewer uses verbal, paraverbal, and non-verbal components of the interview response to evaluate interview performance. In contrast to in-person interviews where the interviewer is explicitly and implicitly processing information, in AVIs the researcher explicitly chooses what predictors to use as inputs in model building (e.g., a researcher can explicitly choose to exclude paraverbal characteristics as predictors when building a ML model). For example, when using an AVI to measure personality, Hickman et al. (2022) measured predictors such as word count, pitch, head pose, and facial action units. While the goal of ML models is to differentiate meaningful from spurious patterns thereby minimizing the contribution of predictors with little predictive validity, the predictors used in the AVI model should still be carefully considered to ensure they are job relevant. For instance, previous studies have found increases in group differences when models include nonverbal behaviors (Booth et al., 2021) and that certain models that use facial images can be less effective on people with darker skin (Buolamwini & Gebru, 2018). In addition, applicants may have negative reactions if they know an AVI is evaluating their facial actions or gestures (Langer et al., 2021). Given these concerns and the intended use of these AVI-CAs in various organizations, the decision was made to only use text-based predictors. The predictors included in the current study are described in the Natural Language Processing (NLP) section within the Study Design and Methods.

Criterion quality is also important. In AVIs, the ML models are optimized to predict the criterion (e.g., human ratings of AVI responses), therefore, if the criterion is incorrectly defined or contains errors, the ML model will reflect this. The underlying quality of the AVI ratings or other criteria (e.g., supervisory ratings of job performance) used to train ML models is a significant determinant of AVI psychometric properties. In the current study, the quality of the human evaluator ratings of AVI responses was foundational both to the quality of ML model predictions and the construct and criterion-related validity of the models. Accordingly, extensive frame-of-reference rater training (Roch et al., 2023) consistent with best practices for enhancing interview structure (Campion et al., 1997) was conducted, along with an examination of the inter-rater reliability of those ratings (see Rater Training and Human Evaluation of Asynchronous Video Interviews in Study Design and Methods) before proceeding with AVI-CA model training.

NLP and ML Techniques

The fields of ML, NLP, and deep learning have seen rapid advancement in recent years. In the generation of text-based features, the first step is transcribing the interview. Once the interview is transcribed, it is necessary to apply NLP to extract information from the transcription (e.g., Chowdhury, 2003). While an exhaustive discussion of different NLP techniques is beyond the scope of this research, we will discuss some of the more prevalent techniques. One technique is bag-of-words where usage frequency is the key data characteristic (El-Din, 2016). While this technique is simple to implement, its drawback is that it does not take context into account. For example, the word “hard” can refer to something having rigid and firm

characteristics, or it can refer to something as being difficult. Bag-of-words techniques would treat both uses of the word as equivalent. Another commonly used NLP technique is sentiment analysis where higher-order meaning is obtained from transcripts using data dictionaries (Serrano-Guerrero et al., 2015). For example, sentiment analysis could be applied to the phrase "I love this restaurant" and, depending on the data dictionary used, it may be coded as "positive" or "happy." While the coding of this statement is simple, more complex statements such as "This show is terrible, but I can't stop watching it" are more difficult to code as it has both positive and negative aspects. A more recent NLP advancement is BERT-based models. BERT refers to Bidirectional Encoder Representations from Transformers and these models consider word context. More specifically, they learn about important information embedded in language, creating dimensional representations of text that are context dependent. For instance, the word "green" in the following three sentences would be represented with different embedding vectors: (a) The company has started a new green initiative to reduce their carbon impact, (b) The family bought a new green colored car, (c) Due to the recent rain, the grass was very green. In this illustration, the vectors for the word "green" used in sentences (b) and (c) refers to a color and would be closer than the vectors for the word "green" used in sentences (a) and (b) where "green" refers both to a color and being environmentally conscious. This results in more accurate language models (Liu et al., 2019). Each of these NLP techniques have their strengths and weaknesses. What technique(s) the researcher chooses to use depends upon theoretical justification along with the performance of the techniques. For instance, sentiment analysis may be useful in determining if a company is viewed positively or negatively in online reviews. In contrast to this useful application of sentiment analysis, in an interview an applicant may talk about a previous negative customer interaction and sentiment analysis may not be able to differentiate the customer being negative versus the applicant being negative, making sentiment analysis less effective. In the current research, both bag-of-words and BERT-based predictors were used as the authors have found both types of data to be incrementally predictive of interview performance (Rottman et al., 2023; see Natural Language Processing (NLP) in Study Design and Methods).

In addition to different transcription and NLP decisions, researchers must make decisions around what ML models to use and how to prevent overfitting their data. Video interview responses can contain thousands of words; therefore, it is important for ML models to distinguish between meaningful and spurious patterns in this big data. Depending on the underlying data, certain models may be more effective at this than others. While it is beyond the scope of this research to compare modeling techniques, models to consider include ridge regression (Hoerl & Kennard, 1970), least absolute shrinkage and selection operator regression (LASSO; Tibshirani, 1996), elastic net (Zou & Hastie, 2005), neural networks (McCulloch & Pitts, 1943), and support vector machines (Cortes & Vapnik, 1995). In addition, researchers are encouraged to use techniques such as cross validation to prevent overfitting (see de Rooij & Weeda, 2020). The current research examined ridge, LASSO, and elastic-net regression models and used 10-fold cross validation in the model building process (see Model Selection and Hyper-parameter Optimization in Study Design and Methods for further details).

One critique commonly leveled against AVIs is that explicit and implicit human biases are built into the models and perpetuated en masse (for an example of human biases being built into ML-based systems see Dastin (2018) and Hunkenschroer (2022)). Different demographic groups can exhibit distinct verbal and paraverbal behaviors; it is possible to identify demographic characteristics such as age, gender, race/ethnicity, and education level through

word usage (Gillick, 2010; Meier et al., 2020; Schler et al., 2006). While the ML models used to build AVIs do not have demographic differences explicitly encoded in them, they could result in subgroup differences if the underlying data used to build the ML model contains subgroup differences or unequal subgroup representation (Köchlig et al., 2020).

While AVIs can demonstrate subgroup differences depending upon the data used to build the ML model, it is also possible that ML models can reduce subgroup differences, although at a cost to convergent validity (Hunkenschroer, 2022; Lukacik et al., 2022). For example, one technique to reduce subgroup differences commonly claimed to be used by hiring technology companies is selectively removing predictors that contribute to subgroup differences (Booth et al., 2021; Raghavan et al., 2020). The current research applied a multi-penalty optimization technique to simultaneously minimize prediction error, degree of overfitting, and subgroup differences, which has been shown to be more effective on interview data than selectively removing predictors (Rottman et al., 2023).

Evaluating Psychometric Properties of AVI-CAs

Reliability

As AVIs represent a form of empirical keying (Hickman et al., 2022), test-retest analysis is considered an ideal index for assessing reliability (Cucina et al., 2019). Test-retest reliability measures the extent to which an individual's scores on an assessment would remain consistent if they were tested more than once. Previous research on test-retest reliability with in-person interviews scored by humans one-year apart have shown lower than desired reliability with r ranging from .26 to .30 depending on whether it was a behavioral, situational, or experience/interest-based interview (Schleicher et al., 2010). In contrast to this, Hickman et al. (2022) found an average test-retest reliability for AVIs designed to measure interviewer-rated measures of personality of $r = .50$ with an average of 15.6 days between the interviews. One reason for the substantially higher test-retest reliability estimates with AVIs may also be partially due to the greater standardization of the administration conditions. For instance, Huffcutt et al. (2013) found that interrater reliability was substantially higher (.74 vs. .44) when interviews were conducted in a single panel session versus separate interviews conducted by each interviewer. With the AVI approach, idiosyncratic interviewer effects are no longer a source of error influencing the evaluation of interviewee performance.

While improvements in standardization may lead to greater stability over time, AVI-CAs are still likely to exhibit lower stability than more traditional psychometric assessments. That is, unlike more traditional psychometric assessments with fixed or constrained response options, interviews represent a more dynamic performance event where the variability in responses is unconstrained (e.g., people may share different events when completing an interview on subsequent occasions). Accordingly, the overall stability of interview responses is likely to be lower than those reported for personality-based measures of the Big Five Factor model (e.g., Costa & McCrae, 1992; e.g., .83 test-retest reliability over a 6-year period for emotional stability). Therefore, the goal of the current study is to examine the test-retest reliability of multiple AVI-CAs. Since there is a lack of research available on the stability of traditional employment interviews (Hickman et al., 2022), there is no clear estimate to compare AVI-CAs against. Nonetheless, synthesizing the available evidence summarized above, we expect test-retest reliability to be comparable or higher than those obtained by Hickman et al. (2022).

Research Question 1: What is the test-retest reliability of AVI-CAs?

Convergent Validity

The validity of AVI-CAs is a critical component in determining if they are appropriate for use in a selection context. While there are several sources of validity (Society for Industrial and Organizational Psychology, 2018) this study focuses on the convergent validity and criterion-related validity of the AVI-CAs.

While traditional test development involves creating assessment content and a scoring rubric, AVIs are unique in that they are created using ML to predict the criterion of interest in a manner similar to empirical keying (Hickman et al., 2022). As such, an examination of convergent validity occurs during model development where the model's objective is to optimize convergent validity.[2] The creation of highly generalizable AVI-CAs presents a unique challenge as it is necessary to use a diverse range of video interviews to maximize generalizability. However, using diverse video interviews, versus more homogenous video interviews, increases the difficulty of the problem the ML model is attempting to solve (i.e., generating a valid prediction of the outcome variable). For example, applicants applying to retail jobs may give homogeneous examples of being willing to learn new skills; in contrast, applicants applying to retail, accounting, and manufacturing jobs are collectively more likely to give heterogeneous examples. While it is likely convergent validity could be improved if job-specific AVI-CAs are built, it is still necessary for more generalizable AVI-CAs to exhibit sufficient levels of convergent validity to support their use in either supplementing or replacing human evaluator ratings.

Research Question 2: What levels of convergent validity do AVI-CAs demonstrate?

AVI-CA Discriminant Relationships

By design, competency modeling is deductive in nature (Campion et al., 2011), first focusing on the outcomes and then identifying the tasks and KSAOs needed to achieve those outcomes. While a strength of the technique is the linking of KSAOs and work objectives, there is no inherent separation between trait-level antecedents and the behaviors underlying each competency. Thus, competency modeling is not first and foremost a construct-pure approach. For instance, a personality trait such as conscientiousness may be an antecedent for the Drive for Results and Initiative, Dependability, and Safety and Compliance Orientation competencies as being goal-directed, organized, and rule following are important characteristics for successful performance across these areas. Each competency can also vary in the degree of overlap among the behaviors specified. For example, while Service Orientation and Developing Others may seem to have less construct overlap, they are both situated in the Work with People domain within the AVI competency framework (see Figure 1). Further, both competencies involve the extent to which one cares about and considers the needs of others. Accordingly, there is commonality in the behavioral antecedents needed to effectively demonstrate each competency.

In addition to the content overlap, both the method (i.e., interview) and the AVI-CA scoring technique may also contribute to higher interrelationships or lower levels of discrimination among different AVI-CAs.[3] In-person interviews have historically demonstrated substantial method variance (Hamdani et al., 2014). Likewise, empirical scoring and ML

[2] In the current study, we operationalize convergent relationships as correlations between human evaluator ratings and algorithmic scores of the same competency area (i.e., between-method comparisons of the same construct).
[3] For discriminant relationships among AVI-CAs, we operationalize them as correlations among AVI-CAs of different competency areas (i.e., within-method comparisons of different constructs).

techniques can often result in high correlations among various constructs due to the high degree of overlap in the predictors included in the models (Park et al., 2015; Simms, 2008). Given this, we wanted to examine the relationships between the competencies, laying the foundation for future construct-related research.[4] While moderate overlap is expected, AVI-CAs should demonstrate higher levels of convergence with ratings or scores on the same competency than ratings or scores on a different competency area to support the distinct value of each AVI-CA measured in a selection context. That is, higher levels of discriminant relationships are indicative of a psychometric instrument with a greater ability to distinguish or differentiate among AVI-CAs.

Research Question 3a: What patterns of discriminant relationships do the AVI-CAs demonstrate?

Research Question 3b: Are the convergent relationships (*Research Question 2*) stronger than the discriminant relationships observed among the AVI-CAs?

Subgroup Differences

To reduce subgroup differences, we applied a multi-penalty optimization technique developed by Rottman et al., (2023). One benefit of this multi-penalty optimization technique is that it can simultaneously be applied across multiple demographic groups (e.g., gender, race/ethnicity, age). Therefore, we wanted to examine the extent to which AVI-CAs demonstrate subgroup differences when multi-penalty optimization is applied during AVI-CA model development.

Research Question 4: To what extent do AVI-CAs demonstrate subgroup differences?

Criterion-related Validity

While it is important that the AVI-CAs demonstrate acceptable levels of convergent validity, it is also important that they relate to key job-relevant organizational criteria such as supervisory ratings of job performance and objective criteria such as sales performance. One current gap in the AVI literature is a lack of criterion-related validity evidence. While the criterion-related validity of an AVI may vary depending on the constructs or KSAOs the assessment is designed to measure, we are unaware of any previous literature that examined their relationship with key organizational criteria. Although Hickman et al.'s (2022) AVI personality assessment was predictive of college GPA, we believe that it is important to examine the relationship of AVI-CAs with organizational performance metrics to ensure viability of AVI-CAs in a selection context. As the goal of the current study was to create highly generalizable AVI-CAs, it is also necessary to examine the extent to which AVI-CAs predict key organizational criteria across jobs, organizations, and industries.

Research Question 5: What levels of criterion-related validity do AVI-CAs demonstrate?

Study Design and Methods

Participants

[4] Since Communication is evaluated across each interview question along with the focal competency area as part of the design of human rater studies, greater degrees of overlap between Communication AVI-CA scores and all other AVI-CA competency model scores is expected.

Sample 1: Asynchronous Video Interviews used to Generate AVI-CAs (Convergent Relationships)

To generate the training set to test convergent validity for each competency area (e.g., Adaptability), it was first necessary to pull competency-specific interview questions from a large database of asynchronous video interviews conducted in the United States. These asynchronous video interviews typically contained five or six interview questions, which is consistent with findings from other research on the average length of asynchronous video interviews (e.g., Dunlop et al., 2022). Generally, three minutes were allotted by the system as the total response time allowed per question and respondents averaged around one minute and forty-five seconds per response. These asynchronous video interviews contained interviews from a variety of jobs, organizations, and industries (see Table 1 for the demographic characteristics of Sample 1). Interview responses came from individuals applying to mostly individual contributor level jobs, ranging from primarily entry-level jobs to some professional-level roles.

To be included in the training set for each competency area, two IO psychologists had to agree that the following conditions were met: the video interview question was well structured (i.e., followed the Situation-Task-Action-Result interview question structure; see Kessler, 2006) and the video interview question would elicit the key behaviors targeted by the competency. Thus, for each competency we selected the most relevant questions completed by applicants that elicited the key behaviors. For each AVI-CA, the training set included more than seven unique interview questions (M = 7.43; SD = 2.44) with a mean of 317.28 interview responses per question (SD = 124.28).[5] Table 2 includes summary statistics on the number of interview questions sampled for each competency area. To avoid the potential for halo or horn effects that could occur when multiple responses from the same interview are presented to raters, a unique or independent set of interviews was included in each AVI-CA training set sample (i.e., there was no overlap in responses from the same recorded interview across competency training set samples). Additionally, Sample 1 is a distinct or non-overlapping set of data from all other samples included in this research study.

Sample 2: Test-retest Reliability Sample

To establish test-retest reliability for each AVI-CA, a sample of applicants who had taken two AVI-CAs with the same competency was gathered from a large database of asynchronous video interviews conducted in the United States. The test-retest sample consisted of applicants who were applying to either the same job multiple times (i.e., subsequent completions after not obtaining the job initially) or who were applying to multiple jobs at either the same organization or a different organization where the AVI-CA included the same specific competency. This sample was similar in composition to the sample used to build the AVI-CAs (Sample 1) in that it was demographically diverse and was primarily composed of individual contributor level jobs. For two of the competencies (Developing Others and Negotiation and Persuasion) there was insufficient data to examine test-retest reliability due to small sample size (< 100 pairs). For the competencies with sufficient sample sizes, the average number of days between time 1 and time 2 was 43.51 days across 181,745 pairs of interview responses. Sample 2 is a distinct or non-overlapping set of data from all other samples included in this research study.

Sample 3: AVI-CA Discriminant Relationships Sample

[5] Since the Communication competency area was rated across interview questions, the area is not included in the summary statistics.

To examine the interrelationship among the different competencies, we gathered a sample of individuals who had taken multiple AVI-CAs which measured distinct competency areas. This sample was similar in composition to Samples 1 and 2 and all interviews were conducted in the United States. This resulted in 48,252 AVI-CAs comprising 34 distinct organizational samples. Sample 3 is a distinct or non-overlapping set of data from all other samples included in this research study.

Samples 4-8: Criterion-Related Validity Samples

To examine criterion-related validity, five United States-based organizational samples (Samples 4-8) were used. Each sample is distinct or non-overlapping with all other research samples. See Table 1 for the self-report demographics across these organizational criterion samples. A description of the job roles, organizations, competencies assessed, validation study design type, and the criterion measures of these five samples are presented in Table 3. In each instance, a job analysis was conducted to determine the best combination of competencies to measure in an AVI-CA. Across all samples, applicants or incumbents took the AVI-CA and then performance metrics were obtained from the organizations.

Samples 6 and 7 provide a good illustration of the information summarized in Table 3 and the range of performance criteria available across the organizational samples. In Sample 6, as part of a concurrent validation study, Customer Service Representative (CSR) agents in a call center for a healthcare company were recruited to complete an AVI-CA designed to measure Communication, Composure, Dependability, Problem Solving, and Team Orientation. A total of 259 incumbents responded to the asynchronous video interview questions and their responses were scored using the appropriate AVI-CAs. All four competency-based structured interview questions were individually scored with the AVI-CA Communication model and then averaged to form an overall Communication score. The scores from all five competency areas were then summed together to form an average interview score. Performance was measured using a three-question customer satisfaction survey completed after an interaction with a CSR either via phone or chat. The survey questions were on a 5-point Likert scale and consisted of questions on how likely the customer is to recommend the company, how satisfied they were with the service received, and how satisfied they were with the resolution provided.

For Sample 7, an asynchronous video interview with a question on Safety Orientation and Compliance was administered as part of the pre-hire selection process for maintenance workers at a large manufacturing organization. The Safety Orientation and Compliance question was then scored using the appropriate AVI-CA model. After employees were on the job for three months, supervisors were given a performance survey with a single question to rate the overall effectiveness of each employee on a 3-point scale ranging from “not meeting expectations” to “exceeding expectations.”

Measures

Behaviorally Anchored Rating Scales (BARS)

Behaviorally anchored ratings scales (BARS) to measure the 15 competencies were developed in conjunction with the competencies, definitions, and interview questions. Specifically, BARS were developed and refined based on the following sources of information: (a) A review of the psychological literature to draw upon behaviors from related constructs,

(b) A review of extant competency frameworks to draw upon BARS from related constructs, (c) Use of critical incidents across job analytic studies to inform construction of behaviorally anchored rating scales, and (d) A review of criterion studies to align BARS behaviors with job performance domains. BARS were first drafted by a team of two IO psychologists from a synthesis of the sources of information noted above. Specifically, the definition and key behaviors for each competency area were reviewed along with the sources of information (a - d above), and then behavioral statements were drafted to represent the low (Novice), middle (Intermediate), and high (Expert) ends of the rating scale. The IO SMEs team worked in tandem on drafting all of the behavioral statements. This process led to a total of 216 behavioral statements across competencies areas, with 12 to 18 statements produced per competency area (M = 14.4, SD = 1.68). Next, focus group sessions were conducted with additional teams of three to five IO psychologists per competency area to ensure behaviors were appropriately specified at each proficiency level. Behavioral statements were revised to improve clarity and avoid redundancy with other statements. Finally, BARS statements were reviewed by a technical writer to ensure clarity and understanding for a broad set of stakeholders.

Samples of applicant responses to actual asynchronous video interview questions were identified for each of the competency domains so that behavioral examples of competencies were assembled to represent each of the five-points on the BARS proficiency level rating scales ranging from 'Novice' to 'Expert.' As an illustration, Table 4 presents the BARS for the Adaptability competency area.

While the full list of interview questions and BARS is proprietary, a sample past behavior structured interview question for Adaptability is: *Tell me about a situation when you had to adapt to a substantial change you encountered while working on a project or school assignment. Please describe the situation, how you adapted to the change, and the outcome.* Past behavioral and situational questions were used to measure these competencies (see Appendix A for a sample interview question from each competency domain).

Demographic Information

Self-reported demographic data were not available from organizations due to data privacy concerns for Samples 1-3. Thus, proprietary demographic prediction models were used to predict the applicant's gender (male and female), race/ethnicity (White, Black, Hispanic, and Asian), and age (under 40 and 40 or older) using video thumbnails.[6] The reported classification accuracy of the model obtained from a prior study was 99% for gender, 87% for race/ethnicity, and 91% for age (McCarthy et al., 2021). It is important to note that the video-based thumbnail data used to predict demographic information was not used in the generation of the audio-based AVI-CAs. For demographic representation, see Table 1 for Samples 1, 4, 5, 7, and 8. Demographic data was not available for Sample 6 because there was neither self-report demographic data nor video data on which to apply the demographic prediction models.

Human Evaluation of Asynchronous Video Interviews

Thirty non-student human evaluators with limited prior experience rating interviews were recruited and hired to rate applicants across asynchronous video interview responses (67% were

[6] The authors would like to acknowledge that the demographic prediction models were only able to broadly classify applicants' gender and race/ethnicity. Demographic prediction models were not able to account for gender being non-dichotomous, the overlap between ethnicity and race (e.g., an applicant has a Hispanic ethnicity and also identifies as Black), Other Races, or Multiracial.

female, 47% were White, 40% were Asian, and 13% were Hispanic). For each competency area rated, evaluators were randomly assigned to rate a subset of video responses from Sample 1 and at least two evaluators rated each video response to produce an average rating score. In addition to rating the specific competency for each interview response, Communication was also rated for each interview response. Evaluators rated an average of almost 2,500 video responses across all 15 competency areas included in the study.

Rater training. Evaluators received frame-of-reference (FOR) training (Roch et al., 2012) to calibrate raters and establish a standard FOR of performance across the proficiency levels for each competency area (see Table 4 for the anchored rating scale and sample behavioral examples). The FOR rater training approach is the most widely researched training approach with meta-analytic studies demonstrating medium (d = .50, Roch et al., 2012) to large (d = .83, Woehr & Huffcutt, 1994) effects on rater accuracy. For the current study, the FOR training consisted of the following elements: (a) an introduction to the structured interviewing approach and its value in reducing raters biases, (b) interview evaluation best practices, (c) the importance of focusing on observable behaviors and avoiding inferences or heuristics, (d) common rater biases and observational errors, (e) defining each competency and its behavioral components, (f) a review of the rating scale and behavior examples at Novice, Intermediate, and Expert proficiency levels, and (g) FOR calibration exercises where raters evaluated applicants' AVI responses to structured interview questions, and received expert feedback on their ratings. The initial introductory training session (elements a through d outlined above) lasted approximately 1.5 to 2 hours and each subsequent FOR calibration training session (elements e through g), conducted separately for each competency area, lasted 1 to 1.5 hours. On average, each FOR calibration session was conducted using 5 applicant interview responses sampled across the range of proficiency levels on the behaviorally anchored rating scale. After evaluators reviewed each candidate response their ratings were recorded and discussed in a group exercise to calibrate raters on their understanding of the competency definitions and BARS and to identify any evidence of cognitive biases and heuristics that could contribute to errors across raters. Additional group FOR calibration sessions were conducted on an ad-hoc basis to ensure continued alignment and avoid drift away from the FOR established. Finally, 1-on-1 training and discussions were provided when raters exhibited biases in their distribution of ratings (e.g., leniency, severity, central tendency), or when they requested feedback on their approach. Collectively, continual training and calibration were provided to ensure a strong understanding of the AVI competency model.

Interview evaluation. To ensure high-quality ratings of asynchronous video interview responses, a high level of structure was applied to the interview content and evaluation approach based on Campion et al.'s (1997) level of structure. The evaluation process was as follows:

1. FOR-based Rater Training was conducted for all evaluators as previously described.
2. To keep ratings focused on the interview response, human evaluators focused solely on the content of interview responses when completing ratings; no ancillary information was available.
3. Human evaluators provided a separate rating for each applicant's response to each interview question using a specific rating scale. To limit the cognitive load of evaluators and promote more accurate judgments, evaluators worked on a

single competency at a time. In doing so, human evaluators watched an applicant's response to the interview question and immediately rated the response before watching and rating a different applicant's response to the same competency.

4. To improve the accuracy of ratings, BARS were used when interviews were rated.
5. Human evaluators were encouraged (but not required) to take notes when making ratings.
6. Ratings for each interview response were statistically averaged across the multiple evaluator ratings to produce a final score. The inter-rater reliability estimates of human evaluators are presented in the Results section.
7. All AVI responses were rated by at least two human evaluators, with the majority being rated by three. Random assignment was conducted with the stipulation that each interview response was rated by at least one male and one female evaluator.
8. Due to the volume of interview responses utilized in this research, having the same human evaluators rate every interview was not feasible. However, each human evaluator rated an average of 2,493 (*Mdn* = 2,470; *SD* = 967) interview question responses across all 15 competency areas.
9. Human evaluators made their interview ratings individually and did not discuss ratings with other evaluators.

Inter-rater reliability. The ability to have a criterion to use in the AVI model training process that has limited measurement error is a critical component or necessary minimum standard to develop algorithms that exhibit useful psychometric properties (e.g., predict job performance). Thus, prior to training any AVI-CAs, inter-rater agreement needed to be quantified to support the aggregation of individual ratings into an average score across raters per competency area. To reduce the potential for idiosyncratic rater effects in the Sample 1 study and for practical purposes (i.e., 30 different raters and different availability of raters), it was necessary to randomly assign multiple evaluators to an interview. It was therefore not possible to have a fully crossed design where all raters overlap on the interviewees they evaluate. This type of study design, where raters and ratees are not fully crossed, is commonly encountered in organizational settings. For example, when gathering multiple supervisor ratings of job performance on individual employees, there is often limited overlap across employees since the manager qualified to evaluate each individual varies. However, as Putka et al. (2008) note, intraclass correlation coefficients (ICCs) are still often calculated even though they produce downwardly biased estimates when assumptions (e.g., raters are nested within ratees, or ratees are fully crossed with raters) are not met. For the current study, given the limited degree of overlap between human raters (i.e., the same two raters did not consistently evaluate the same interviewees and the number of raters varied per interview), our study design is considered an ill-structured measurement design (ISMD; Putka et al., 2008). That is, there was a random assignment of 30 different evaluators to rate each interview question, producing a sparse matrix of overlap between individual evaluators. Accordingly, the most appropriate metric to use for calculating inter-rater reliability is $G(q,k)$, which was developed by Putka et al. (2008). The $G(q,k)$ reliability estimator was specifically designed for ISMDs where there is a violation of a core assumption of the ICC (Shrout & Fleiss, 1979) - that there is independence of residual errors across raters. To improve reliability estimations under the condition of non-independence

of residual errors across raters, the $G(q,k)$ formula explicitly accounts for the amount of overlap between the sets of raters who rate each interviewee. When rater studies are fully crossed (e.g., all ratees are fully crossed with raters), the $G(q,k)$ formula is equivalent to the ICC(C,k) formula (McGraw & Wong, 1996).

Table 5 presents estimates of $G(q,k)$ across all 15 AVI-CA areas. The average $G(q,k)$ estimate across competencies was .66, with a range from .62 to .70. This is slightly higher than the average single rater reliability reported in Campion et al. (2016) and consistent with inter-rater reliability levels reported by Hickman et al. (2022).

Model Development Analytic Strategy

A rich set of data points from applicant verbal responses is generated in an average 15- to 30-minute video interview. To generate the AVI-CAs, these data points were extracted from the video responses and were then used as the predictors in the ML modeling process with the criterion being the human evaluator ratings of the target competency. The process for creating each AVI-CA model is outlined in detail below.

Natural Language Processing (NLP)

To build the AVI-CAs, it was first necessary to take audio files recorded from the video interview responses and transform them into data that can then be used in a model. To do this, speech-to-text transcription occurred where raw audio was analyzed and then converted to text using a third-party deep learning model (Rev AI, 2021), resulting in binary bag-of-words predictors. In addition to the bag-of-word predictors, a BERT-based approach to NLP called RoBERTa (Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach; Liu et al., 2019) was applied to the transcripts. The RoBERTa model used to generate the AVI-CA models was first fine-tuned on video interview transcripts by taking the standard RoBERTa model as developed by Liu et al., (2019) and training it to predict human evaluator ratings of Communication. This allowed the RoBERTa model to better learn about how words are used in an interview context. Word embeddings from the final layer of this transformer-based NLP model were used to generate 768 RoBERTa predictors. Both bag-of-words and RoBERTa predictors were used in all AVI-CA models. See Table 6 for the number of predictors in each AVI-CA model.

Model Selection and Hyper-parameter Optimization

Before a ML model can be trained and validated, a process of model selection and hyper-parameter[7] optimization occurs (e.g., Bergstra & Bengio, 2012). This process works by iterating through different ML model types and hyper-parameter values (e.g., lambda) and then examining the error on a portion of the data held out from training to assess the applicability of those settings to the training problem. It is important to note that none of the weighting of predictor

[7] Hyper-parameters are higher-level model settings such as learning rate and regularization strength, that govern how a model learns for a given problem and set of data. For Ridge Regression, for instance, an example of a hyper-parameter fixed prior to model training is lambda (λ). Lambda is used to restrict the magnitude of allowable regression weights of predictors in a model. The lambda hyper-parameter essentially balances under and overfitting of a model. When lambda is equal to 0, then a Ridge regression model produces the same coefficients as a simple linear regression. When the lambda value is high, the regression coefficients are reduced toward zero to avoid overfitting.

variables is determined in this process. This process is merely used to determine the best way to approach the training problem or learn appropriate inferences from the data.

For model selection and hyper-parameter optimization, we examined the predictive validity of three regression-based models (ridge, lasso, and elastic-net regression) along with multiple model parameters using 10-fold cross validation (for more information about cross validation see Kohavi, 1995). In 10–fold cross validation, the total sample is split into 90% training sample and 10% test (validation) sample 10 times. A model is then trained using the data from the training sample and then used to create predictions for the test sample. This process is repeated a total of ten times. Therefore, ten separate models were each trained on 9/10th of the data set and the resulting model was then applied on a unique $1/10^{th}$ of the data. The model predicted scores and the human evaluator ratings for each test sample were combined into one dataset and model accuracy was calculated by correlating model predicted scores with human evaluator ratings. Across all AVI-CAs, ridge regression models had higher or equivalent levels of predictive validity than elastic net and lasso models. Given this, we decided to use ridge regression models for all AVI-CAs.

Final Model Building and Multi-penalty Optimization

Once a model type (i.e., ridge regression) and hyper-parameters were selected, those settings were used for training and validating the final model on the entire dataset. While demographic characteristics were not explicitly used as predictors, there is the risk that proxies for demographic group differences in the form of predictors (e.g., word embeddings; Bolukbasi et al., 2016) or outcome/dependent variables (e.g., human evaluator ratings in the current study) can lead ML models to reproduce or amplify group differences encoded in training data. To minimize the potential for subgroup differences, a technique called multi-penalty optimization was used to build the final model (Rottman et al., 2023). In this technique, a hyperparameter that penalizes models with high subgroup differences (β) was added to the algorithm. This results in ridge regression models that simultaneously minimize prediction error, degree of overfitting, and subgroup differences. $\beta \geq 0$ is a tunable parameter such that higher levels of β produce models with smaller subgroup differences and smaller predictive validity, while lower levels of β produce models with larger subgroup differences and larger predictive validity. Two IO psychologists examined the diversity-validity tradeoff of the various β hyperparameters and decided on the final β hyperparameter to balance the tradeoff. That is, the researchers opted for final models that reduced subgroup differences to a trivial to small level with a minimal reduction in convergent validity (for further reading about this technique, see Rottman et al., 2023). Final model hyperparameters can be found in Table 6.

Transparency and Openness

We describe our data and analyses in the study and adhered to the Journal of Applied Psychology methodological checklist. A complete list of interview questions, interview transcripts, interview videos, and performance measures are not available due to their proprietary nature. De-identified predictors, applicant demographics, and performance metrics are available upon request. The study design and analyses were not pre-registered.

The majority of analyses and the ML model building process was conducted using Python 3.8.10 (Van Rossum & Drake, 2009) and the following packages: Scikit-learn 0.21.1 (Pedregosa et al., 2011), pandas 1.2.3 (Reback et al., 2021), NumPy 1.19.5 (Harris et al., 2020), and scipy 1.6.0 (Virtanen et al., 2020). Calculations of the same weighted average correlations

and construction of meta-analytic confidence intervals were conducted using R 3.6.0 (R Core Team, 2022) using the psychometric package (Fletcher, 2022).

Results

Means and standard deviations for both the average human evaluator ratings and predicted model scores can be found in Table 7 .

Test-Retest Reliability of AVI-CAs

To determine the test-retest reliability of the AVI-CAs (*Research Question 1*), data was analyzed from actual candidates that had taken two instances of an AVI-CA with the same competency (Sample 2). The average correlation across all competencies was .72, indicating that the AVI-CAs demonstrated adequate test-retest reliability. The average test-retest correlation for the Communication AVI-CA was considerably higher than other competencies ($\bar{r}$ = .82). This is to be expected since the Communication AVI-CA was scored across all structured interview questions (M = 5.18; SD = 1.18; n = 181,610) and is therefore more likely to exhibit higher stability. See Table 8 for test-retest reliability estimates across 12 of the 15 AVI-CAs.

Convergent Validity Evidence for AVI-CAs

The convergent validity of the AVI-CA model scores and human evaluator ratings was examined (Sample 1, *Research Question 2*). The AVI-CAs effectively replicated human evaluator ratings with Multiple *R*s ranging from .55 to .74 with a sample weighted average convergent validity correlation of .66 (see Table 7 for competency-specific results).

AVI-CA Patterns of Discriminant Relationships

The interrelationships among the AVI-CAs (*Research Question 3a*) was examined by evaluating the sample weighted average correlations across 34 organizational samples where AVI-CAs were launched in a selection context[8] (Sample 3). A full correlation matrix of the sample weighted average intercorrelations ($\bar{r}$) among the individual AVI-CAs, as well as the sample weighted average correlation for each competency with all other competencies (overall $\bar{r}$) was produced as part of this analysis (see Table 9). It is important to note that certain combinations of competencies were not included in a single interview assessment across the organizational samples, therefore, we were unable to examine interrelationships for these competency pairs. In examining the interrelationships, Communication was highly related with the other competencies (overall $\bar{r}$ = .74, 95% CI [.73, .75]). The majority of competencies (with the exception of Communication) had moderate interrelationships ranging from overall $\bar{r}_{Team\ Orientation}$ = .48 (95% CI [.46, .48]) to overall $\bar{r}_{Composure}$ = .61 (95% CI [.59, .62]).

For *Research Question 3b*, the convergent validity correlations obtained from Sample 1 (Table 7, *Research Question 2*) were compared to the overall sample weighted average discriminant correlation coefficients reported in Table 9 (*Research Question 3a*). AVI-CA scores for a given competency area had significantly higher correlations (p < .05) with expert human

[8] The respective AVI-CA algorithmic scoring developed in Sample 1 was simply applied to the AVI-CAs in Sample 3 to score them. No new model training or development occurred.

evaluator ratings on the same competency area (Sample 1) than they did with AVI-CA scores with different competency areas (Sample 3) for 13 out of 15 competency areas. No significant difference in correlations was found for Negotiation and Persuasion (z = 1.40, p = 0.08). Communication was the only area that had significantly lower correlations with expert human evaluator ratings on the same competency area than it did with AVI-CAs scores with different competency areas (z = -25.74, p < 0.01).

The pattern of weaker discriminant relationships between Communication and the other competencies, the moderately high correlations found among the other competencies, and the generally higher levels of convergent validity between human and algorithmic evaluations of the same competency in Sample 1, suggest that while the AVI-CAs are distinct in what they are measuring, there is likely a high common method effect.

Subgroup Differences for Human Evaluations and AVI-CAs

Tables 10 and 11 present Cohen's *d* demographic subgroup differences for average human evaluator ratings across all 15 competency areas (see Online Supplement Table S2 for full means and standard deviations). There were minimal subgroup differences in human evaluator ratings across all competency areas (weighted average Cohen's d = -.02; n = 70,445)[9] which is consistent with the magnitude of subgroups typically found with structured interview ratings (e.g., Huffcutt & Roth, 1998; Sackett et al., 2022). Subgroup differences in AVI-CA scores (*Research Question 4*) were examined across all AVI-CAs in Sample 1 and across mean AVI-CAs scores in Samples 4-8 (criterion-related validity samples). Table 8 also presents subgroup differences across all 15 AVI-CAs for Sample 1. For Sample 1, gender, age, and race/ethnicity subgroup differences for AVI-CA scores had Cohen's *d* values that ranged from -.14 to .06 (weighted average Cohen's d = -.01, n = 732,567), indicating minimal subgroup differences. For all AVI-CAs, subgroup differences were comparable or lower than the subgroup differences found in the trained human evaluator ratings.

Criterion Samples. For Samples 4, 5, 7, and 8[10], gender and race/ethnicity subgroup differences for AVI-CA scores had Cohen's *d* values that ranged from -.20 to .10 (weighted average Cohen's d = -.04, n = 37,199) and confidence intervals which included 0 in twelve of the fourteen comparisons (see Table 12 and Online Supplement Table S3). Similar to the results from Sample 1, these results indicated trivial to small subgroup differences in the applicant populations where the AVI-CAs were used in a selection context.

Criterion-Related Validity of AVI-CAs

Using data from five samples (Samples 4-8), the relationship between the mean AVI-CA score and performance was examined to obtain an estimate of criterion-related validity (*Research Question 5)*. Since limited information was available on how each AVI-CA was used along with other selection decision tools (e.g., an in-person interview, a cognitive ability test, or an assessment center), no corrections for restriction of range, attenuation due to criterion unreliability, or any other corrections for artifacts were performed. This provides a more conservative estimate of the criterion-related validity for the AVI-CAs. Computation of sample mean weighted correlations across criterion study samples allowed us to compare the

[9] For ease of interpretation, all Cohen's *d* effect sizes were calculated using White, Male, and Under 40 years old as the reference group for their respective demographic category.

[10] Subgroup differences for Sample 6 are not included since no self-report demographic data was available and no video data was available to apply demographic prediction models.

uncorrected criterion-related validity of AVI-CAs with uncorrected meta-analytic estimates for human ratings of structured interviews (Sackett et al., 2022).

Criterion-related validity in the five studies ranged from .20 in the maintenance worker sample (Sample 7) to .27 (Sample 4) in the call center worker sample (see Table 3[11] for criterion-related validities across each sample). An average, sample weighted uncorrected meta-analytic criterion-related validity estimate of $\bar{r}$ = .24 was found across the five studies (n = 1,124). This indicates that the AVI-CAs can predict performance across a wide variety of jobs and industries. The validity estimate in the current study ($\bar{r}$ = .24) is lower but in a comparable range to those of human-rated structured interviews in prior meta-analytic results (Sackett et al., 2022 - uncorrected $\bar{r}$ = .32).

Discussion

This study adds to the field of personnel selection by examining the psychometric properties of AVI-CAs which were developed with the intent of being highly generalizable. Results demonstrated that AVI-CAs can be developed with adequate levels of reliability, convergent validity, moderate to high interrelationships, and minimal levels of demographic subgroup differences. Equally or more important, these AVI-CAs significantly predicted job performance in five organizations.

The magnitude of convergent validity obtained across the AVI-CAs ranged from .55 to.74, with a weighted average convergent validity coefficient of .66. Further, 12 of 15 AVI-CAs had higher convergent validity effect sizes than the effect sizes found in Campion et al.'s (2016) study using essay responses. This improvement is likely the result of the richness of the AVI response data (versus typed essay-based responses) and improvements in NLP. Further, the magnitude of convergent validities obtained in the current study with AVI-CAs were also substantially higher than those obtained from AVI personality assessments created using interviewer-reported personality ($\bar{r}$ = .37; Hickman et al., 2022). Several principal differences between the current AVI-CA study and Hickman et al.'s study are the constructs examined, the level of analysis and evaluation methods, and the type of study design (i.e., field in the present study versus lab in the former study). Thus, differences in the magnitude of convergent validities obtained across these two studies may be due to both construct and study design factors. More specifically, an important distinction between the current study and Hickman et al.'s (2022) approach is the level of analysis employed, with the current study using the question level of analysis while the former used the interview level of analysis to build AVI scoring models. Further, the use of BARS in the current study versus self-report (Likert-type responses to behavioral statements) and interviewer-judged personality (also using a Likert-type scale to rate interviewee personality) in the Hickman et al. (2022) study is likely also to account for the higher levels of convergence observed in the current study. The use of BARS primes raters to focus directly on the observable behaviors within a specific interview question. This enhances rater accuracy (Roch et al. 2012) and simplifies the prediction task for ML models versus broad personality judgments made across an interview.

Overall, convergent correlations (Sample 1) tended to be higher than the average correlations with other competencies (Sample 3), except for correlations with Communication. This provides initial evidence that the AVI-CA scores are able to differentiate among competencies. However, the collective evidence is mixed, given that Communication

[11] For full correlation matrices on criterion Samples 4 through 8, see Online Supplement Tables S4 – S8.

consistently exhibited high or higher interrelationships with different AVI-CAs than it did with human evaluator ratings on the same construct. While moderate to high intercorrelations among AVI-CAs measuring different competencies were exhibited, initial evidence suggests that the models appear to exhibit stronger patterns of discriminant validity than studies conducted with human evaluators (e.g., van Iddekinge et al., 2004). In prior research, the construct validity evidence - expected patterns of larger magnitudes in convergent validity (i.e., average correlations among ratings or scores of the same construct) compared to discriminant validity (i.e., average correlations among ratings or scores of different constructs) - is mixed with some studies finding lower convergent than discriminant correlations (Facteau et al., 2000; Huffcutt et al. 2001; van Iddekinge et al., 2004), others finding moderate levels of higher convergent patterns (Motowidlo et al. 1992), and some finding very small differences, but higher convergent validity (Conway & Peneno, 1999).

In a study employing a multi-trait, multi-method approach, van Iddekinge et al. (2004) found that the mean convergent validities across two interviews designed to assess the same constructs - Interpersonal Skills, Conscientiousness, and Stress Management - were weaker than the discriminant relationships among the constructs (i.e., correlations among ratings for interview questions targeting different constructs). It is important to note that the interview sessions in the van Iddekinge et al. (2004) study were conducted live and as such there is the potential for more idiosyncratic effects (e.g., how the interviewer conducts the session) that could increase the interviewer effect, and consequently inflate the magnitude of divergent validity effect sizes. In summary, while estimates from the current study are more favorable in terms of initial evidence of discriminant patterns than estimates for studies on human structured interview ratings, they are generally consistent with the mixed results found in the structured interview literature to date.

Consistent with research discussed by Bleidorn and Hopwood (2019), correlations in the current study may be inflated due to overlap in the response characteristics across all the AVI-CAs. As Park et al. (2015) aptly discuss, because the central goal was to create AVI-CAs with high predictive accuracy (i.e., convergence with trained human evaluator ratings), we included all word-based (i.e., words and NLP features) predictors available within a competency training set when building each of the AVI-CAs. As a result, an AVI-CA for any given competency domain can have substantial overlap with the predictors of an AVI-CA for a different competency domain merely because the same word or NLP category appears in responses to different competency-based questions. Ultimately, this overlap in features may increase correlations. Future research should examine the impact of removing predictors shared across models on patterns of convergent and discriminant validity, as removing these features would likely increase discriminant validity, but may decrease convergent validity. Future research should also explore whether fine-tuning NLP models specifically on a single competency area could further improve discriminant validity without sacrificing convergent validity.

Ratings of Communication were completed for each question along with the focal competency across each competency area training set (e.g., for questions designed to measure Adaptability, a rating for Communication occurred at the same time a rating for Adaptability was made). This was a practical design decision as it allowed us to reduce the length of the interview by measuring the Communication competency across all interview questions (i.e., it was not necessary to have a specific Communication-focused question). While Communication is intended to be an independent, albeit related, construct, it is important to recognize that it is inextricably linked to performance on the focal competency area measured. For instance, if one cannot exhibit effective Communication behaviors in an interview (e.g., the ‘Delivers Clear &

Concise Message' key behavior), more limited behavioral information will be available to evaluate the focal competency, likely resulting in a lower focal competency score. Consequently, if someone demonstrates high levels of Communication in the interview, they are more likely to have presented a clear and structured response that provides an opportunity to demonstrate higher levels of the focal competency. Thus, this introduction of common method variance (Podsakoff et al., 2003) likely also contributes to the high degree of overlap between Communication and the focal competency areas evaluated in Sample 1. Future research may benefit from a focus on more explicitly separating constructs and methods by independently training algorithms on questions designed to directly measure Communication.

The AVI-CAs showed good stability over time (*Research Question 4*), with average test-retest reliability results of $r = .72$ obtained across AVI-CAs being higher than those obtained AVI-PAs (.50 average; Hickman et al., 2022). Finally, subgroup differences on mean predicted model scores by protected classes revealed trivial to small differences, consistent with research on structured interview ratings (Huffcutt & Roth, 1998). Thus, the models are consistent and fair with respect to their scoring of applicants from different demographic backgrounds.

In addition to the convergent validity and competency discriminant relationships examined with Research Questions 2 and 3a - 3b, a direct examination of the criterion-related validity of AVI-CAs in predicting job performance was conducted using predictive and concurrent criterion validation studies (*Research Question 5*). Results across jobs and industries indicate significant criterion-related validity, ranging from .20 to .27. The uncorrected sample-weighted criterion-related validity estimate in the current study ($\bar{r} = .24$) is lower but in a comparable range to those of human-rated structured interviews in prior meta-analytic results (Sackett et al., 2022 - uncorrected $\bar{r} = .32$). It is also important to note that the current estimate for AVI-CAs is based on 5 studies while the Sackett et al.'s (2022) meta-analysis included 105 effect sizes from primary structured interviews studies. Thus, with this body of research in its infancy, there is much more evidence to accumulate, and these estimates will continue to be refined as more data becomes available on AVI-CAs. Nonetheless, initial estimates suggest that interviews evaluated by AVI-CAs demonstrated similar, albeit slightly lower, predictive validity than human-rated interviews, while demonstrating desirable psychometric properties (test-retest reliability, convergent and criterion-related validity) and minimal subgroup differences consistent with research on structured interview ratings (Huffcutt & Roth, 1998). Given the cost advantages of ML scored interviews, they present a compelling alternative for organizations to consider over other early-stage selection tools (e.g., a recruiter phone screen).

Methodological and Applied Contributions

While the AVI-CAs were built to measure competencies as demonstrated in AVIs, the methodology detailed in this paper can be applied to a wide variety of selection scenarios. For example, while this research built broadly generalizable AVI-CAs using AVIs from multiple organizations, it is also possible to apply the methodology outlined in this research to build organizational-specific AVI-CAs or even job-specific AVI-CAs (discussed further below). The methodology outlined in the present research provides organizations, consultants, and researchers a framework for developing AVIs that can be tailored to meet specific organizational goals and objectives.

Another substantive contribution of this study was the examination of subgroup differences in AVI-CAs. Minimal to small subgroup differences were found across training and

criterion samples. It is important to note that this study employed several strategies to decrease the potential for subgroup differences including: ensuring that the data going into the models (predictors) was demographically diverse, extensive rater training and the application of modeling techniques which decreased subgroup differences (i.e., multi-penalty optimization). Thus, AI developers and stakeholders of such systems are encouraged to adopt and adhere to professional testing standards (e.g., Society for Industrial Organizational Psychology, 2018, 2022) and emerging AI-scored selection tool audit frameworks to determine the extent to which AI-based assessments are valid and fair for specific use cases (Landers & Behrend, 2022). Therefore, while there is increased concern and greater legal regulations around the use of AVIs, it would be premature and shortsighted to categorically ban or limit the use of AVI-CAs and other carefully developed AI-based decision tools as they have the potential to have lower levels of subgroup differences than human evaluators, which we found in the present study.

Deploying these models at an enterprise scale can provide large cost and process efficiency savings compared to traditional human resource structured interview programs. For example, where average human evaluators take 30 minutes to rate each AVI and screen 25,000 applicants a year, approximately 12,500 annual hours of human evaluator time will be saved if AVI-CAs scores are used to screen applicants.

Potential Limitations and Future Research

The current AVI-CAs were built using asynchronous video interviews from various companies, industries, and job families. While this was done intentionally to improve generalizability of the AVI-CAs, future research should examine the incremental validity of organizational- or job-specific models. In addition, future research should examine the methodology implemented in the current study with related selection contexts (e.g., phone interviews, resumes, written work samples, recorded sales pitches, language-centric assessment center exercises).

The current AVI-CAs were built using word-based features only (i.e., words and NLP features) to focus on the most job relevant behaviors that can be measured in an AVI selection context. We explicitly excluded paraverbal (e.g., pitch, cadence, tone), and nonverbal (e.g., facial expressions) variables. Future research could benefit from examining the incremental validity of paraverbal and nonverbal features in ML models that include words and NLP features. However, with increasing legal concerns and the potential for visual-based data to be less accurate for certain demographic groups (Buolamwini & Gebru, 2018), a substantial body of evidence on the incremental validity and negligible impact on subgroup differences of such features needs to be accumulated before they are used in a selection context.

Although the current study allowed for estimates of criterion-related validity of AVI-CAs to be compared against meta-analytic estimates of traditional structured interview ratings, data did not exist to examine the direct relationship between human evaluator ratings and job performance since independent human evaluator ratings were not collected when the AVI-CAs were deployed in Samples 4-8 for operational use. Future research would benefit from a true multi-trait, multi-method examination (Campbell & Fiske, 1959) where raters are blind to AVI-CA scores and their independent human evaluator AVI ratings and/or other measures of these competencies are directly compared to job performance outcomes along with AVI-CA scores. Most importantly, such a study would allow for an examination of convergent and discriminant

validity across methods, and criterion-related validity of each method, expanding our understanding of the AVI-CA nomological network.

To understand the extent to which criterion-related validity varies between asynchronous and live structured interviews, future research would benefit from an experimental design where candidates are randomly assigned to take either a synchronous or asynchronous interview and evaluated by both humans and algorithms. Such a study would allow for examination of both the effects of synchronicity and evaluation method on criterion-related validity. For example, candidates could be more (or less) engaged in structured interviews when interacting with an interviewer versus recording responses asynchronously, creating a stronger (or weaker) situation for individual differences to be activated and expressed. In an AVI-CA, while the absence of probing is good from a standardization of administration conditions perspective, allowing for effective interviewer follow-up probing questions may have the potential to elicit more job-relevant behaviors than an AVI format where candidates do not receive any immediate cues to elaborate on an incomplete response (Levashina et al., 2014).

One critique commonly leveled against ML models is their "black box" nature (Gonzalez et al., 2019). While some may argue that explainability is unnecessary and that the efficacy of ML models should be based on their usefulness over time (Norvig, 2017), future research should apply advances in the field of explainable artificial intelligence to better understand automated video assessments (see Adadi & Berrada, 2018, for a review on explainable AI). Understanding how a model works can provide numerous benefits. From an AVI development perspective, explainability can provide evidence on the extent to which the AVIs are working as intended. For example, if predictors deemed to be important in the model are theoretically aligned with the construct the model is intended to measure, the researcher can be more confident the model is properly specified. In contrast, if predictors deemed to be important in the model are not theoretically aligned with the construct the model is intending to measure, it serves as important feedback to guide the researcher toward potential model modifications. From an applicant perspective, where applicants are often distrustful of AI, explainability could facilitate the generation of a feedback report allowing the applicant to understand why they received the assessment score that they did, ultimately improving trust. Finally, from a researcher-perspective, understanding ML models could help refine existing theories about the constructs measured in the AVI. For example, in a Problem Solving AVI, explainability could highlight key behaviors that are important to the construct of Problem Solving that may have not previously been identified in the initial conceptualization of the construction thereby advancing theory.

Similar to the impact of NLP design choices, choices made in the model building process could have produced different results. For example, a smaller multi-penalty optimization hyperparameter (β) could have been applied which would have increased the convergent validity but would also have resulted in higher subgroup differences. This is just an example of how ML decisions can have important ramifications. In addition, the fields of NLP and ML are rapidly advancing. Given this, we believe that the current levels of predictive validity found with the AVI-CAs are, in fact, the lower limits of predictive validity that we could expect to see as these fields further advance. A partnership between the fields of Industrial/Organizational Psychology and ML would be beneficial to ensure that the most appropriate NLP and ML techniques are applied, and psychometric properties of such techniques are appropriately studied. One potential new application of ML is the use of language-agnostic BERT modeling (Feng et al., 2020), which captures dimensional representations of text (i.e., contextualization of words or phrases) across languages. While considerable research is needed to investigate the viability of this

approach across different contexts (i.e., most prominently culture and language differences), language-agnostic BERT modeling would allow for the development of AVI scoring algorithms across languages. Future research should examine the effectiveness of this solution and explore the generalizability of the current AVI-CAs across different countries, languages, and cultures.

Conclusion

The present study demonstrated the strong psychometric properties of AVI-CAs designed for use across multiple jobs, organizations, and industries. While future research is needed to better understand and improve AVIs, this research is an important step in demonstrating their effectiveness across a wide variety of contexts, providing researchers, organizations, and consultants new tools to screen applicants both effectively and efficiently with minimal subgroup differences.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138 52160. 10.1109/ACCESS.2018.2870052

Arikan, S., Erktin, E., & Pesen, M. (2020). Development and validation of a STEM competencies assessment framework. *International Journal of Science and Mathematics Education, 1-24*. https://doi.org/10.1007/s10763-020-10132-3

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Artificial Intelligence Video Interview Act, ILCS § HB2557 (2020). https://www.ilga.gov/legislation/fulltext.asp?DocName=&SessionId=108&GA=101&DocTypeId=HB&DocNum=2557&GAID=15&LegID=118664&SpecSess=&Session=

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology*, *90*(6), 1185–1203. https://doi.org/10.1037/0021-9010.90.6.1185

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305. https://dl.acm.org/doi/10.5555/2188385.2188395

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. https://doi.org/10.1177/1088868318772990

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Booth, B. M., Hickman, L., Subbaraj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021). Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction,* 268–277. https://doi.org/10.1145/3462244.3479897

Brenner, F. S., Ortner, T. M., & Fay, D. (2016). Asynchronous video interviewing as a new technology in personnel selection: The applicant's point of view. *Frontiers in Psychology*, *7*, 863. https://doi.org/10.3389/fpsyg.2016.00863

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research*, *81*, 77–91.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64*, 225-262. https://doi.org/10.1111/j.1744-6570.2010.01207.x

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*(3), 655–702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*, 958–975. https://doi.org/10.1037/apl0000108

Carroll, A., & McCrackin, J. (1998). The competent use of competency-based strategies for selection and development. *Performance Improvement Quarterly, 11*, 45-63.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology, 9*(3), 621–640. https://doi.org/10.1017/iop.2016.6

Chapman, D. S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment*, *11*(2–3), 113-120. https://doi.org/10.1111/1468-2389.00234

Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M., & Yoon, S.-Y. (2016). Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, 161–168. https://doi.org/10.1145/2993148.2993203

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, E. M. (2018). Automated video interview judgment on a large-sized corpus collected online. *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, 504–509. https://doi.org/10.1109/ACII.2017.8273646

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Conway, J.M., Peneno, G.M. Comparing Structured Interview Question Types: Construct Validity and Applicant Reactions. *Journal of Business and Psychology, 13*, 485–506 (1999). https://doi.org/10.1023/A:1022914803347

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cucina, J. M., Vasilopoulos, N. L., Su, C., Busciglio, H. H., Cozma, I., DeCostanza, A. H., Martin, N. R., & Shaw, M. N. (2019). The effects of empirical keying of personality measures on faking and criterion-related validity. *Journal of Business and Psychology*, *34*(3), 337–356. https://link.springer.com/article/10.1007/s10869-018-9544-y

Dano, E. B. (2019, October). A validated systems engineering competency methodology and functional/domain competency assessment tool. In *2019 International Symposium on Systems Engineering (ISSE)* (pp. 1-7). IEEE

Dastin J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Dunlop, P. D., Holtrop, D., & Wee, S. (2022). How asynchronous video interviews are used in practice: A study of an Australian-based AVI vendor. *International Journal of Selection and Assessment*. https://doi.org/10.1111/ijsa.12372

El-Din, D. M. (2016). Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *7*(1). https://doi.org/10.14569/IJACSA.2016.070134

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, *1*, 216–243.

Facteau, J.D., Facteau, C. L., Jackson, K.A. & Becton, J.B. (2000, April). Do structured interviews measure KSAs? An investigation of the construct validity of interview ratings in two organizations. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, News Orleans, LA.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*

Fletcher, T. D. (2022). Psychometric: Applied Psychometric Theory. https://CRAN.R-project.org/package=psychometric.

Gillick, D. (2010) Can conversational word usage be used to predict speaker demographics? *Proceedings Interspeech 2010*, 1381-1384. 10.21437/Interspeech.2010-421

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). "Where's the IO?" Artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, *5*(3), 5. https://doi.org/10.25035/pad.2019.03.005

Gorman, C. A., Robinson, J., & Gamble, J. S. (2018). An investigation into the validity of asynchronous web-based video employment-interview ratings. *Consulting Psychology Journal: Practice and Research, 70*(2), 129–146. https://doi.org/10.1037/cpb0000102

Green, P. (1999). *Building robust competencies: Linking human resource systems to organizational strategies*. San Francisco, CA: Jossey-Bass.

Hamdani, M. R., Valcea, S., & Buckley, M. R. (2014). The relentless pursuit of construct validity in the design of employment interviews. *Human Resource Management Review*, *24*(2), 160–176. https://doi.org/10.1016/j.hrmr.2013.07.002

Harris, C. R., Millman, K. J., van der Walt, S. J., . . . Oliphant, T. E. (2020), Array programming with NumPy. *Nature*, *585*, 357–362.1038/s41586-020-2649-2

Hartwell, C. J., Johnson, C. D., & Posthuma, R. A. (2019). Are we asking the right questions? A validity comparison of structured interview question types. *Journal of Business Research, 100*, 122–129. https://doi.org/10.1016/j.jbusres.2019.03.026

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8), 1323–1351. https://doi.org/10.1037/apl0000695

Hickman, L., Tay, L., & Woo, S. E. (2019). Validity evidence for off-the-shelf language based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. *Personnel Assessment and Decisions*, *5*(3), 12–20. https://doi.org/10.25035/pad.2019.03.003

HireVue. (2017). Unilever finds top talent faster with HireVue assessments. https://cdn.featuredcustomers.com/CustomerCaseStudy.document/hirevue_unilever_138410.pdf

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82. 10.1080/00401706.1970.10488635

Hogan, R., Chamorro-Premuzic, T., & Kaiser, R. B. (2013). Employability and career success: Bridging the gap between theory and reality. *Industrial and Organizational Psychology*, *6*(1), 3–16. https://doi.org/10.1111/iops.12001

Howard, P. F., Liang, Z., Leggat, S., & Karimi, L. (2018). Validation of a management competency assessment tool for health service managers. *Journal of Health Organization and Management, 32*, 113-134.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897–913. https://doi.org/10.1037/0021-9010.86.5.897

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, *21*(3), 264–276. https://doi.org/10.1111/ijsa.12036

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology*, *83*(2), 179–189. https://doi.org/10.1037/0021-9010.83.2.179

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., Degroot, T. G., & Jones, C. (2001). Comparison of situational and behavior description interview questions for higher-level positions. Personnel Psychology, 54(3), 619–644. https://doi.org/10.1111/j.1744-6570.2001.tb00225.x

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 1–31. https://doi.org/10.1007/s10551-022-05049-6

Hunter, J.E. and Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings (2nd ed.).* Thousand Oaks: Sage Publications.

Kessler, R. (2006). *Competency-based interviews*. Pompton Plains, NJ: Career Press.

Köchling, A., Riazy, S., Wehner, M., & Simbeck, K. (2020). Highly accurate, but still discriminatory: A fairness evaluation of algorithmic video analysis. *Academy of Management Proceedings, 2020*(1), 10510. https://doi.org/10.5465/AMBPP.2020.13339abstract

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence, 14*, 1137–1145.

Kochanski J. (1997). Competency-based management. *Training and Development, 51*, 41–44.

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist.* Advance online publication. http://dx.doi.org/10.1037/amp0000972

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment, 29*, 154–169. https://doi.org/10.1111/ijsa.12325

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and

third-party observers. *Computers in Human Behavior*, *123*, 106878. https://doi.org/10.1016/j.chb.2021.106878

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241-293. https://doi.org/10.1111/peps.12052

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arxiv:1907.11692

Lukacik, E.- R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), 100789. https://doi.org/10.1016/j.hrmr.2020.100789

McCarthy, J. M., Truxillo, D. M., Bauer, T. N., Erdogan, B., Shao, Y., Wang, M., Liff, J., & Gardner, C. (2021). Distressed and distracted by COVID-19 during high-stakes virtual interviews: The role of job interview anxiety on performance and reactions. *Journal of Applied Psychology*, *106*(8), 11031117. https://doi.org/10.1037/apl0000943

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *American Psychologist, 28*, 1–14. https://doi.org/10.1037/h0034092

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Meier, T., Boyd, R. L., Mehl, M. R., Milek, A., Pennebaker, J. W., Martin, M., Wolf, M., & Horn, A. B. (2020). Stereotyping in the digital age: Male language is “ingenious”, female language is “beautiful” - and popular. *PLoS ONE*, *15*(12). https://doi.org/10.1371/journal.pone.0243637

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., & Vaughan, M. J. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77(5)*, 571–587. https://doi.org/10.1037/0021-9010.77.5.571

Nolan, P. (1998). Competencies drive decision making. *Nursing Management, 29(3)*, 27-29.

Norvig, P. (2017). *Google’s approach to artificial intelligence and machine learning*. UNSW. https://www.youtube.com/watch?v=oD5Ug6uO0j8

Oliphant, G. C., Hansen, K., Oliphant, B. J. (2008). A review of a telephone-administered behavior-based interview technique. *Business and Professional Communication Quarterly, 71*(3), 383–386. https://doi.org/10.1177/1080569908321429

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J. Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET.* American Psychological Association. https://doi.org/10.1037/10313-000

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

R Core Team (2022). *R: A language and environment for statistical computing* (Version 3.6.0). R Foundation for Statistical Computing. https://www.R-project.org/.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Rasipuram, S., Rao, P., & Jayagopi, D. B., (2016). Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: A systematic study. *Proceedings of the 18th ACM International Conference on Multimodal Interaction,* 370–317. https://doi.org/10.1145/2993148.2993183

Reback, J., McKinney, W., . . . & Dong, K. (2021). pandas-dev/pandas: Pandas. *Zenodo*. 10.5281/zenodo.4572994

Rev AI. (2021). *The world's most accurate API for AI- and human-generated transcripts* (Version 1) [Computer software]. https://www.rev.ai/

Roch, S.G., Woehr, D.J., Mishra, V. and Kieszczynska, U. (2012), Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, 85: 370-395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Rottman, C., Gardner, C., Liff, J., Mondragon, N., & Zuloaga, L. (2023). New strategies for addressing the diversity-validity dilemma with big data. *Journal of Applied Psychology*, *108*(9), 1425-144. https://doi.org/10.1037/apl0001084

Rubinstein, P. (2020, November 5). *Asynchronous video interviews: The tools you need to succeed*. BBC. https://www.bbc.com/worklife/article/20201102-asynchronous-video-interviews-the-tools-you-need-to-succeed

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000994

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology, 95*(4), 603–617. https://doi.org/10.1037/a0018920

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *American Association for Artificial Intelligence Spring Symposium on Computational Approaches for Analyzing Weblogs*, *6*, 199–205.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38. http://dx.doi.org/10.1016/j.ins.2015.03.040

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

Society for Industrial and Organizational Psychology, (2018). *Principles for the Validation and Use of Personnel Selection Procedures* (Fifth). American Psychological Association. https://doi.org/10.1017/iop.2018.195

Society for Industrial and Organizational Psychology (2022). *SIOP statement on the use of artificial intelligence (AI) for hiring: Guidance on the effective use of AI-based assessments.* https://www.siop.org/Portals/84/docs/SIOP%20Statement%20on%20the%20Use%20of%20Artificial%20Intelligence.pdf?ver=mSGVRY-z_wR5iIuE2NWQPQ%3d%3d

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. *Personnel Assessment and Decisions*, *7*(2), 1–22. https://doi.org/10.25035/pad.2021.02.001

Torres, E. N., & Gregory, A. (2018). Hiring managers' evaluations of asynchronous video interviews: The role of candidate competencies, aesthetics, and resume placement. *International Journal of Hospitality Management*, 75, 86–93. https://doi.org/10.1016/j.ijhm.2018.03.011

Trevethan, R. 2016. Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, 117.

Ullah, Z., Lajis, A., Jamjoom, M., Altalhi, A. H., Shah, J., & Saleem, F. (2019). A rule-based method for cognitive competency assessment in computer programming using Bloom's taxonomy. *IEEE Access, 7,* 64663-64675.

van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., Jr., & Attenweiler, W. J. (2004). What Do Structured Selection Interviews Really Measure? The Construct Validity of Behavior Description Interviews. *Human Performance*, *17*(1), 71–93. https://doi.org/10.1207/S15327043HUP1701_4

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T. E., . . . & SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Woehr, D.J. and Huffcutt, A.I. (1994), Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67: 189-205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Table 1

Applicant Demographics Across Study Samples

Demographic Group	*N* (Sample %)				
	Sample 1	Sample 4	Sample 5	Sample 7	Sample 8
Gender					
Male	13,443 (48%)	796 (56%)	596 (18%)	2,960 (73%)	1,446 (12%)
Female	14,380 (52%)	633 (44%)	2,777 (82%)	1,080 (27%)	10,672 (88%)
Age					
Under 40	23,966 (86%)	–	–	–	–
Over 40	3,857 (14%)	–	–	–	–
Ethnicity					
White	9,936 (36%)	960 (70%)	889 (27%)	1,406 (11%)	2,940 (26%)
Black	4,770 (17%)	264 (19%)	1,932 (58%)	2,271 (18%)	6,030 (54%)
Hispanic	9,316 (33%)	143 (10%)	503 (15%)	145 (1%)	2,034 (18%)
Asian	3,800 (14%)	–	–	–	148 (1%)
Two or more races	–	–	–	8,567 (69%)	–
Level					
Non-Manager	20,333 (76%)	–	–	–	–
Manager	6,577 (24%)	–	–	–	–

Table 2

Sample 1 Descriptive Statistics on Training Set Interview Questions

Competency Label	Training Sample Interview Question Statistics				*N* of Training Sample
	Total Questions	Mean Number of Responses per Question	*SD*	Median	
Adaptability	12	235.42	202.89	231.50	2,829
Communication	–	–	–	–	29,948
Compassion	10	193.30	101.95	216.00	1,934
Composure	6	386.83	300.00	363.41	2,401
Coordination of People and Resources	4	446.25	83.03	452.50	1,786
Dependability	7	364.86	330.00	168.48	2,554
Developing Others	10	181.50	108.00	178.50	1,818
Drive for Results/Initiative	9	311.89	206.85	220.00	2,812
Negotiation & Persuasion	7	198.00	187.70	148.00	1,386
Problem Solving	4	640.50	319.48	724.50	2,562
Relationship Building	8	237.88	50.51	236.00	2,110
Safety & Compliance Orientation	9	246.22	15.54	212.00	2,221
Service Orientation	8	275.13	156.23	199.50	2,201
Team Orientation	5	380.60	250.68	381.00	1,903
Willingness to Learn	5	343.60	125.66	398.00	1,718

Table 3

Samples 4 – 7 Criterion Validation Study Results on Competency Models

Sample Name	Overview	Competency Models Used	Validation Study Design	Criteria	*n*	Observed Criterion Validity (*r*)[a]
Sample 4	Call Center Insurance Sales for a large direct-to-consumer health insurance organization.	Drive for Results & Initiative	Predictive	Mean Revenue Marketing Spend[b]	228	0.27**
Sample 5	Retail Call Center for American sporting goods store chain.	Adaptability, Communication, Composure, Dependability, and Service Orientation	Predictive	Quality Assurance[c]	165	0.23**
Sample 6	Insurance Call Center Customer Service	Dependability, Composure, Problem Solving, Communication, and Team Orientation	Concurrent	Customer Satisfaction Survey Score[d]	259	0.26**
Sample 7	Maintenance Worker in a Large Manufacturing Organization	Safety Orientation & Compliance	Predictive	Supervisory Performance Rating[e]	292	0.20**
Sample 8	Insurance Call Center Customer Service	Composure & Problem Solving	Predictive	Supervisory Performance Rating[e]	180	0.23**

[a]The Criterion Related Validity is calculated by correlating the average score across all interview competency algorithms measured in the sample with the criterion measure. [b]The amount of revenue generated compared to the marketing expenses needed to generate that revenue. [c]Quality Assurance is a weighted average of supervisor evaluations of agent performance, adherence to standards, and customer satisfaction survey scores. [d]Customer Satisfaction Survey is a weighted average of customer survey responses after an interaction with a Call Center Representative, measured on a 5-point Likert scale, and consisting of questions on how likely the customer is to recommend the company, how satisfied they were with the service received, and how satisfied they were with the resolution provided. [e]90-days post-hire supervisors were asked to rate their employees' overall level of job performance on a 3-point scale ranging from *Not Meeting Expectations* to *Exceeding Expectations*.
$^{**}p < .01$.

Table 4

Behaviorally Anchored Rating Scale for Adaptability

Adaptability					
This competency refers to the ability to shift or change opinions, actions, or behaviors. Those ranking high in this competency can successfully adjust when faced with multiple demands, shifting priorities, rapid change, or ambiguity.					
Key Behaviors	**Novice (1)**	**Developing (2)**	**Intermediate (3)**	**Advanced (4)**	**Expert (5)**
	Proficiency Level Rating Guidelines:				
	Candidate is unlikely to be successful in situations requiring this competency.	Candidate is likely to demonstrate this competency in simple situations or in a limited capacity.	Candidate is likely to demonstrate this competency well, but may need assistance in more difficult situations.	Candidate is likely to demonstrate this competency effectively in moderate to complex situations.	Candidate is likely to demonstrate this competency with extreme effectiveness in moderate to complex situations.
	Behavioral Examples at Novice, Intermediate, and Expert Proficiency Levels:				
Sees the Positive in Change	Reacts negatively to the change; is concerned about all of the extra effort they will have to put forth while adjusting.		May be somewhat reluctant to accept the change at first, but is able to see positive aspects after internalizing the change.		Views the change as a positive challenge or opportunity for learning and growth.
Seeks to Understand Change	Does not seek information to understand the need for the change.		Understands the change is needed as a part of job requirements.		Possesses a detailed understanding of the change, and its benefits to the company as well as their own role.
Adjusts Behavior to Accommodate Change	Does not effectively adjust their behavior as required to meet the demands of the situation.		Adjusts their behavior within a reasonable timeframe to meet demands of a moderate to difficult change.		Quickly modifies behavior to meet demands of a difficult, complex, or time intensive change.
Drives the Change	Requires supervisor or manager to make the necessary changes.		Requires minimal guidance.		Inspires others to embrace the change while modeling appropriate behaviors.

Table 5

Sample 1 Interrater Reliability across Ratings

Competency Label	Total Ratings[a]	$G(q,k)$	$\widehat{k}$[b]
Adaptability	7,141	0.68	2.52
Communication	59,859	0.63	2.23
Compassion	3,812	0.66	2.08
Dependability	4,970	0.63	2.16
Composure	5,776	0.67	2.43
Coordination of People & Resources	3,318	0.62	2.01
Developing Others	3,651	0.62	2.22
Drive for Results/Initiative	6,687	0.67	2.42
Negotiation & Persuasion	2,236	0.65	2.19
Problem Solving	4,503	0.67	2.29
Relationship Building	4,070	0.64	2.08
Safety & Compliance Orientation	4,105	0.70	2.15
Service Orientation	3,998	0.63	2.08
Team Orientation	5,334	0.70	2.80
Willingness to Learn	4,211	0.67	2.54
Weighted Average Inter-rater Reliability across Competency Areas	123,671	0.66	2.28

Note. $G(q,k)$ for the Communication competency area was estimated by first computing $G(q,k)$ for each set of Communication ratings within a focal competency area (e.g., Communication ratings on Adaptability interview questions), and then a weighted average of all 14 areas was computed to estimate overall $G(q,k)$ for Communication.

[a]Total Ratings represents the total number of ratings completed across all raters. [b] $\widehat{k}$ represents the harmonic mean number of raters/evaluators per ratee/interviewee.

Table 6

Sample 1 Model Descriptives

Competency Label	*N* of Predictors[a]	λ	β
Adaptability	9,175	1,000.0	1.0
Communication	23,847	1,000.0	7.0
Compassion	6,454	1,000.0	5.0
Dependability	8,103	1,000.0	0.1
Coordination of People & Resources	8,336	1,000.0	1.0
Developing Others	6,804	1,000.0	5.0
Drive for Results/Initiative	8,459	1,000.0	1.0
Negotiation & Persuasion	6,512	1,000.0	1.0
Composure	7,555	1,000.0	0.1
Problem Solving	5,218	1,000.0	0.5
Relationship Building	7,325	1,000.0	1.0
Safety & Compliance Orientation	8,055	1,000.0	1.0
Service Orientation	6,572	1,000.0	0.5
Team Orientation	4,154	1,000.0	0.5
Willingness to Learn	6,296	1,000.0	0.5

Note. λ (lambda) represents the hyper-parameter used to restrict the magnitude of allowable regression weights of predictors in a model, which balances under and overfitting. β is the multi-optimization hyperparameter penalty which penalizes models with high subgroup differences. [a]Of the total number of predictors, 768 are RoBERTa-based and the remaining number are individual words.

Table 7

Sample 1 Training Sample Sizes, Means and Standard Deviations for Human Evaluator and Predicted Competency Model Scores, and Convergent Validity

Competency Label	*N* of Training Sample	Human Evaluator Ratings		Predicted Competency Model Scores		Model Performance (Convergent Validity)	
		M	*SD*	*M*	*SD*	Multiple *R*	Multiple *R* 95% C.I.
Adaptability	2,829	3.22	0.69	3.22	0.38	0.65	0.63 – 0.67
Communication	29,948	3.04	0.73	3.04	0.47	0.66	0.65 – 0.67
Compassion	1,934	3.06	0.83	3.06	0.59	0.74	0.72 – 0.76
Composure	2,401	3.02	0.76	3.02	0.49	0.68	0.65 – 0.70
Coordination of People and Resources	1,786	3.01	0.87	3.00	0.56	0.70	0.68 – 0.73
Dependability	2,554	3.09	0.79	3.09	0.50	0.67	0.64 – 0.69
Developing Others	1,818	2.99	0.85	2.99	0.49	0.65	0.62 – 0.68
Drive for Results/Initiative	2,812	3.03	0.76	3.02	0.49	0.69	0.67 – 0.71
Negotiation & Persuasion	1,386	2.89	0.88	2.89	0.44	0.58	0.54 – 0.61
Problem Solving	2,562	2.76	0.94	2.76	0.57	0.65	0.63 – 0.67
Relationship Building	2,110	3.06	0.80	3.06	0.49	0.70	0.68 – 0.72
Safety & Compliance Orientation	2,221	3.10	0.80	3.10	0.53	0.70	0.68 – 0.72
Service Orientation	2,201	3.14	0.73	3.14	0.48	0.68	0.65 – 0.70
Team Orientation	1,903	2.90	0.74	2.90	0.34	0.55	0.52 – 0.58
Willingness to Learn	1,718	3.21	0.70	3.21	0.33	0.58	0.55 – 0.61

Note. Multiple *R* is based on out-of-sample algorithm scores from 10 k-fold runs; 95% confidence interval constructed based on the standard deviation of *R* across all 10 k-fold runs.

Table 8

Sample 2 Test-Retest Reliability for Competency Model Scores

Competency Area	r^a	*n*	Mean Days	Median Days	*SD* Days	Min Days	Max Days
Communication	0.82	37,135	43.39	11	66.58	0	501
Compassion	0.71	135	14.06	7	21.25	0	125
Adaptability	0.74	12,895	45.04	9	67.90	0	398
Dependability	0.65	38,491	43.46	11	66.38	0	501
Team Orientation	0.68	35,548	40.43	10	62.47	0	398
Willingness to Learn	0.69	19,013	49.18	13	71.97	0	501
Composure	0.7	15,109	48.47	11	72.80	0	501
Problem Solving	0.69	15,082	46.71	10	72.70	0	501
Drive for Results/Initiative	0.80	3,773	28.81	11	45.81	0	398
Relationship Building	0.75	1,433	29.27	8	41.48	0	232
Service Orientation	0.67	2,296	27.00	14	34.50	0	232
Safety & Compliance Orientation	0.51	393	21.30	8	29.98	0	146
Coordination of People & Resources	0.57	442	22.62	8	31.55	0	153

[a]Test-retest Pearson correlation coefficient between assessment scores completed at Time 1 and Time 2.

Table 9

Sample 3 AVI-CA Discriminant Relationships

Competency	Statistics	Overall $\bar{r}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Adaptability	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.53 [.52, .54]														
	n [*k*]	151,920 [27]														
2. Communication	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.74 [.73, .75]	.72 [.70, .74]													
	n [*k*]	241,781 [34]	33,359 [27]													
3. Compassion	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.56 [.55, .57]	.57 [.56, .58]	.75 [.73, .76]												
	n [*k*]	42,120 [8]	6,972 [4]	8,448 [8]												
4. Composure	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.61 [.59, .62]	.61 [.57, .65]	.83 [.81, .85]	.54 [.41, .66]											
	n [*k*]	57,649 [15]	10,367 [12]	13,332 [15]	927 [2]											
5. Coordination of People and Resources	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.59 [.57, .61]	.63 [.62, .64]	.80 [.79, .82]	.58 [.50, .66]	.64 [.60, .69]										
	n [*k*]	12,576 [5]	1,130 [2]	2,926 [5]	526 [2]	871 [3]										
6. Dependability	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.55 [.53, .56]	.52 [.50, .55]	.77 [.76, .79]	.51 [.50, .53]	.62 [.56, .68]	.59 [.50, .67]									
	n [*k*]	157,637 [29]	25,307 [23]	35,144 [29]	8,351 [6]	3,896 [8]	1,638 [3]									
7. Developing Others	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.61 [.58, .64]	.62 [.58, .66]	.80 [.78, .83]	.56 [-]	.65 [.63, .67]	.63 [.59, .68]	.65 [-]								
	n [*k*]	10,199 [5]	744 [2]	2,312 [5]	454 [1]	837 [3]	1,594 [3]	452 [1]								
8. Drive for Results/Initiative	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.53 [.51, .55]	.56 [.53, .58]	.73 [.70, .76]	.54 [-]	.62 [.61, .64]	.63 [.55, .70]	.55 [.50, .59]	-							
	n [*k*]	75,081 [17]	12,215 [13]	15,674 [17]	73 [1]	3,265 [3]	179 [2]	12,278 [16]	-							
9. Negotiation & Persuasion	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.55 [.53, .57]	.63 [.59, .66]	.76 [.72, .80]	-	-	.48 [-]	.55 [.53, .56]	-	.56 [.52, .59]						
	n [*k*]	3,659 [3]	108 [2]	997 [3]	-	-	70 [1]	925 [3]	-	876 [3]						
10. Problem Solving	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.51 [.50, .53]	.49 [.46, .53]	.67 [.63, .71]	.60 [.56, .64]	.46 [.42, .50]	.54 [.49,.59]	.52 [.48, .56]	.53 [.46, .61]	.47 [.42, .51]	.54 [-]					
	n [*k*]	102,294 [17]	16,250 [14]	21,469 [17]	6,944 [3]	1,412 [5]	1,182 [4]	19,508 [14]	1,253 [3]	9,233 [11]	156 [1]					
11. Relationship Building	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.56 [.55, .58]	.57 [.54, .60]	.79 [.77, .80]	.55 [.52, .58]	.64 [.59, .70]	.61 [.57, .65]	.59 [.56, .62]	.64 [.56, .71]	.59 [.55, .63]	.62 [.57, .66]	.51 [.47, .55]				
	n [*k*]	101,897 [25]	17,064 [18]	22,642 [25]	1,507 [5]	4,984 [10]	2,831 [4]	19,329 [21]	2,207 [3]	10,262 [10]	394 [2]	9,691 [8]				
12. Safety & Compliance Orientation	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.53 [.52, .54]	.48 [.46, .50]	.74 [.72,.76]	.54 [.52, .56]	.58 [.55, .60]	.53 [.47, .58]	.54 [.52, .57]	.50 [.48, .53]	.47 [.42, .52]	-	.53 [.49, .60]	.52 [.50, .54]			
	n [*k*]	105,599 [19]	15,715 [12]	23,692 [19]	7,886 [5]	8,055 [9]	1,173 [3]	17,975 [16]	1,174 [3]	5,166 [6]	-	9,114 [6]	7,504 [10]			
13. Service Orientation	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.58 [.57, .60]	.55 [.53, .58]	.79 [.78, .80]	.59 [.56, .61]	.64 [.62, .67]	.64 [.51, .77]	.58 [.54, .61]	.69 [.65, .73]	.61 [.61, .61]	.52 [.44, .60]	.60 [.56, .64]	.65 [.61, .69]	.57 [.55, .60]		
	n [*k*]	85,031 [13]	14,808 [10]	17,828 [13]	8,079 [6]	9,542 [10]	99 [2]	11,135 [11]	742 [2]	961 [4]	356 [2]	8,205 [4]	3,429 [9]	14,737 [11]		
14. Team Orientation	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.48 [.46, .50]	.45 [.39, .51]	.61 [.56, .66]	-	.59 [.54, .64]	.63 [-]	.48 [.42, .53]	.63 [-]	.41 [.36, .47]	-	.43 [.37, .50]	.47 [.40, .53]	.50 [.46, .54]	.52 [.48,.57]	
	n [*k*]	61,365 [19]	10,994 [15]	14,649 [19]	-	6,201 [8]	29 [1]	8,919 [15]	34 [1]	4,461 [9]	-	5,626 [10]	4,707 [6]	5,493 [8]	5,351 [9]	
15. Willingness to Learn	$\bar{r}$ [$\bar{r}$ 95% *CI*]	.52 [.51, .53]	.51 [.49, .54]	.69 [.67, .72]	.48 [.37, .59]	.60 [.57, .64]	.53 [.51, .55]	.55 [.52, .58]	-	.49 [.46, .53]	.54 [.53, .55]	.48 [.43, .53]	.54 [.51, .57]	.51 [.47,.55]	.57 [.53, .60]	.45 [.40,.51]
	n [*k*]	124,781 [29]	20,246 [21]	29,309 [29]	401 [3]	5,690 [10]	1,254 [2]	26324 [28]	-	12,519 [16]	774 [2]	13,720 [13]	17,992 [19]	11,607 [14]	4,704 [8]	9,550 [15]

Note. Overall $\bar{r}$ =the sample weighed average correlation across all competencies, k = number of unique organizational samples; $\bar{r}$ = the sample weighted average correlation; A 95% confidence interval is constructed based on the uncorrected mean correlation, accounting for sampling error only (see Hunter & Schmidt, 2004); Missing values indicates both competencies were not measured in the same interview.

Table 10

Sample 1 Ethnicity Subgroup Differences on Human Evaluator and Algorithm Competency Ratings

Competency	Statistics	Human Evaluator Ratings				Competency Model Scores			
		White[a]	Asian	Black	Hispanic	White[a]	Asian	Black	Hispanic
Adaptability	Cohen's *d* [*d* 95% C.I.]	–	-0.05 [-0.17 - 0.08]	0.10 [-0.02 - 0.23]	-0.09 [-0.18 - 0.01]	–	-0.05 [-0.08 – -0.03]	-0.06 [-0.08 – -0.04]	-0.04 [-0.06 – -0.01]
	n	1,449	298	297	557	23,952	9,727	12,093	6,895
Communication	Cohen's *d* [*d* 95% C.I.]	–	-0.18 [-0.23 – -0.14]	-0.03 [-0.06 – 0.00]	-0.08 [-0.11 – -0.05]	–	-0.04 [-0.06 – -0.01]	-0.05 [-0.07 – -0.03]	-0.05 [-0.07 – -0.03]
	n	11,099	2,345	5,241	8,639	37,883	8,567	17,075	18,379
Compassion	Cohen's *d* [*d* 95% C.I.]	–	-0.13 [-0.36 – 0.11]	-0.03 [-0.16 – 0.11]	-0.13 [-0.24 – -0.03]	–	-0.11 [-0.23 – 0.01]	-0.03 [-0.08 – 0.01]	-0.05 [-0.09 – -0.00]
	n	643	78	321	748	4,850	282	3,121	2,921
Composure	Cohen's *d* [*d* 95% C.I.]	–	-0.03 [-0.18 – 0.13]	0.05 [-0.07 – 0.17]	0.04 [-0.06 – 0.14]	–	-0.08 [-0.12 – -0.05]	-0.13 [-0.16 – -0.11]	-0.06 [-0.09 – -0.04]
	n	845	189	411	732	23,965	3,631	11,266	8,446
Coordination of People and Resources	Cohen's *d* [*d* 95% C.I.]	–	-0.05 [-0.21 – -0.11]	0.05 [-0.11 - 0.21]	-0.07 [-0.18 – 0.04]	–	-0.05 [-0.15 – 0.04]	-0.06 [-0.15 – 0.02]	-0.14 [-0.21 – -0.07]
	n	541	208	221	706	2,883	491	694	1,006
Dependability	Cohen's *d* [*d* 95% C.I.]	–	-0.05 [-0.19 – 0.09]	-0.06 [-0.17 – 0.04]	0.04 [-0.06 – 0.14]	–	0.06 [0.04 – 0.08]	0.02 [-0.01 – 0.05]	-0.08 [-0.10 – -0.05]
	n	864	252	561	640	11,067	36,607	14,375	20,341
Developing Others	Cohen's *d* [*d* 95% C.I.]	–	-0.09 [-0.32 – 0.15]	-0.13 [-0.28 – 0.02]	-0.06 [-0.17 – 0.04]	–	-0.13 [-0.24 – -0.02]	-0.05 [-0.16 – 0.06]	-0.12 [-0.20 – -0.04]
	n	614	81	235	777	1,589	392	431	912

[a]For ease of interpretation White is the reference group across all comparisons. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Table 10 con't

Sample 1 Ethnicity Subgroup Differences on Human Evaluator and Algorithm Competency Ratings

Competency	Statistics	Human Evaluator Ratings				Competency Model Scores			
		White[a]	Asian	Black	Hispanic	White[a]	Asian	Black	Hispanic
Drive for Results and Initiative	Cohen's *d* [*d* 95% C.I.]	–	-0.10 [-0.22 – 0.02]	-0.04 [-0.18 – 0.10]	0.01 [-0.08 – 0.11]	–	-0.03 [-0.06 – 0.00]	-0.02 [-0.05 – 0.00]	-0.07 [-0.09 – -0.04]
	n	1,265	346	243	659	23,598	5,288	7,606	9,844
Negotiation and Persuasion	Cohen's *d* [*d* 95% C.I.]	–	0.02 [-0.15 – 0.20]	0.01 [-0.16 – 0.19]	0.04 [-0.13 – 0.20]	–	-0.05 [-0.13 – 0.03]	-0.03 [-0.08 – 0.01]	-0.08 [-0.12 – -0.03]
	n	813	152	151	179	7,325	628	2,918	2,522
Problem Solving	Cohen's *d* [*d* 95% C.I.]	–	0.00 [-0.23 – 0.23]	0.04 [-0.05 – 0.14]	0.04 [-0.08 – 0.16]	–	-0.02 [-0.05 – -0.00]	-0.04 [-0.06 – -0.02]	-0.00 [-0.02 – 0.03]
	n	1,078	78	713	374	26,036	9,922	13,290	11,301
Relationship Building	Cohen's *d* [*d* 95% C.I.]	–	-0.19 [-0.37 – -0.01]	-0.01[-0.14 – 0.12]	-0.01 [-0.11 – 0.10]	–	-0.01 [-0.05 – 0.03]	-0.06 [-0.10 – -0.02]	-0.08 [-0.11 – -0.04]
	n	649	147	363	821	12,917	2,791	3,495	4,070
Safety and Compliance Orientation	Cohen's *d* [*d* 95% C.I.]	–	-0.22 [-0.38 – -0.06]	-0.07 [-0.20 – 0.07]	-0.09 [-0.20 – 0.03]	–	-0.04 [-0.09 – -0.01]	-0.06 [-0.09 – -0.02]	0.02 [-0.02 – -0.05]
	n	423	248	472	922	11,397	1,668	4,614	4,271
Service Orientation	Cohen's *d* [*d* 95% C.I.]	–	-0.19 [-0.39 – 0.02]	0.07 [-0.13 - 0.13]	-0.14 [-0.25 – -0.02]	–	-0.02 [-0.05 – 0.01]	-0.02 [-0.05 – 0.00]	-0.02 [-0.04 – 0.01]
	n	581	110	415	955	15,640	4,071	8,229	7,560

[a]For ease of interpretation White is the reference group across all comparisons. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Table 10 con't

Sample 1 Ethnicity Subgroup Differences on Human Evaluator and Algorithm Competency Ratings

Competency	Statistics	Human Evaluator Ratings				Competency Model Scores			
		White[a]	Asian	Black	Hispanic	White[a]	Asian	Black	Hispanic
Team Orientation	Cohen's *d* [*d* 95% C.I.]	–	0.03 [-0.20 – 0.26]	0.15 [0.03 – 0.27]	-0.19 [-0.33 – -0.05]	–	-0.04 [-0.06 – -0.01]	-0.06 [-0.08 – -0.03]	-0.03 [-0.06 – -0.01]
	n	748	83	406	397	21,093	10,512	9,897	8,831
Willingness to Learn	Cohen's *d* [*d* 95% C.I.]	–	0.00 [-0.22 – 0.22]	0.17 [0.06 – 0.29]	0.10 [-0.05 – 0.26]	–	-0.02 [-0.05 – 0.00]	-0.06 [-0.08 – -0.04]	-0.03[-0.06 – -0.01]
	n	741	92	493	202	29,283	8,135	13,632	10,686

[a]For ease of interpretation White is the reference group across all comparisons. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Table 11

Sample 1 Gender and Age Subgroup Differences on Human Evaluator and Algorithm Competency Ratings

Competency	Statistics	Human Evaluator Ratings				Competency Model Scores			
		Male[a]	Female	Under 40 years old[a]	Age 40 or older	Male[a]	Female	Under 40[a] years old	Age 40 or older
Adaptability	Cohen's *d* [*d* 95% C.I.]	–	-0.03 [-0.04 - 0.10]	–	-0.06 [-0.16 – 0.05]	–	0.00 [-0.02 – 0.02]	–	-0.00 [-0.03 – 0.02]
	n	1,530	1,071	2,179	422	27,651	25,018	45,944	6,725
Communication	Cohen's *d* [*d* 95% C.I.]	–	0.02 [0.00 – 0.04]	–	-0.01 [-0.05 – 0.02]	–	0.01 [-0.01 – 0.02]	–	-0.00 [-0.02 – 0.02]
	n	12,569	14,758	23,583	3,744	38,577	43,333	45,944	6,725
Compassion	Cohen's *d* [*d* 95% C.I.]	–	0.08 [-0.03 - 0.20]	–	-0.04 [-0.19 – 0.11]	–	0.00 [-0.05 – 0.05]	–	-0.01 [-0.05 – 0.04]
	n	360	1,430	1,601	189	2,167	9,008	8,960	2,215
Composure	Cohen's *d* [*d* 95% C.I.]	–	0.01 [-0.11 – 0.13]	–	-0.03 [-0.15 – 0.09]	–	0.04 [0.02 – 0.06]	–	-0.01 [-0.05 – -0.01]
	n	1,224	953	1,858	319	21,092	26,221	38,440	8,873
Coordination of People and Resources	Cohen's *d* [*d* 95% C.I.]	–	-0.09 [-0.19 – 0.01]	–	0.05 [-0.08 – 0.18]	–	0.02 [-0.04 - 0.08]	–	-0.04 [-0.10 – 0.02]
	n	947	729	1,396	280	2,167	2,907	1,620	3,454
Dependability	Cohen's *d* [*d* 95% C.I.]	–	0.05 [-0.03 – 0.13]	–	-0.11 [-0.25 – 0.03]	–	0.05 [-0.03 – 0.06]	–	-0.02 [-0.04 – 0.00]
	n	1,196	1,121	2,110	207	36,981	45,416	70,199	12,198
Developing Others	Cohen's *d* [*d* 95% C.I.]	–	-0.07 [-0.16 – 0.03]	–	-0.02 [-0.16 – 0.12]	–	0.01[-0.06 – 0.07]	–	-0.00 [-0.06 – 0.06]
	n	728	979	1,479	228	1,495	1,829	1,014	2,310

[a]Reference groups for ease of interpretation across all comparisons are White and Under 40 years old. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Table 11 con't

Sample 1 Gender and Age Subgroup Differences on Human Evaluator and Algorithm Competency Ratings

Competency	Statistics	Human Evaluator Ratings				Competency Model Scores			
		Male[a]	Female	Under 40 years old[a]	Age 40 or older	Male[a]	Female	Under 40 years old[a]	Age 40 or older
Drive for Results and Initiative	Cohen's *d* [*d* 95% C.I.]	–	-0.15 [-0.23 – -0.07]	–	-0.04 [-0.06 – 0.14]	–	0.01[-0.00 – 0.03]	–	-0.03 [-0.05 – 0.00]
	n	1,413	1,100	2,055	458	23,741	22,598	39,413	6,926
Negotiation and Persuasion	Cohen's *d* [*d* 95% C.I.]	–	-0.02 [-0.10 – 0.13]	–	-0.17 [-0.32 – -0.01]	–	0.05 [-0.02 – -0.08]	–	-0.03 [-0.07 – 0.01]
	n	611	684	1,103	192	6,689	6,704	10,053	3,340
Problem Solving	Cohen's *d* [*d* 95% C.I.]	–	0.02 [-0.08 – 0.12]	–	-0.10 [-0.21 – 0.01]	–	0.02 [0.01 – 0.04]	–	-0.01 [-0.03 – 0.02]
	n	512	1,731	1,888	355	29,435	31,119	53,185	7,369
Relationship Building	Cohen's *d* [*d* 95% C.I.]	–	0.02 [-0.07 – 0.10]	–	-0.15 [-0.26 – -0.04]	–	0.03 [-0.01 – -0.06]	–	-0.01 [-0.04 – 0.03]
	n	903	1,077	1,758	147	14,246	9,029	18,844	4,431
Safety and Compliance Orientation	Cohen's *d* [*d* 95% C.I.]	–	-0.04 [-0.13 – -0.05]	–	0.03 [-0.10 – 0.16]	–	-0.00 [-0.03 – 0.03]	–	-0.01 [-0.05 – 0.03]
	n	1,274	792	1,816	250	10,774	11,179	18,485	3,468
Service Orientation	Cohen's *d* [*d* 95% C.I.]	–	-0.07 [-0.16 – 0.01]	–	0.15 [0.02 – 0.28]	–	0.04 [0.02 – 0.06]	–	-0.01 [-0.04 – 0.02]
	n	820	1,241	1,795	266	17,099	18,406	29,924	5,581

[a]Reference groups for ease of interpretation across all comparisons are White and Under 40 years old. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Table 11 con't

Sample 1 Gender and Age Subgroup Differences on Human Evaluator and Algorithm Competency Ratings

Competency	Statistics	Human Evaluator Ratings				Competency Model Scores			
		Male[a]	Female	Under 40[a] years old	Age 40 or older	Male[a]	Female	Under 40[a] years old	Age 40 or older
Team Orientation	Cohen's *d* [*d* 95% C.I.]	–	0.01 [-0.06 – 0.07]	–	0.02 [-0.13 – 0.16]	–	0.03 [0.01 – 0.05]	–	-0.03 [-0.06 – 0.01]
	n	1,495	1,829	1,424	211	29,165	21,171	46,116	4,220
Willingness to Learn	Cohen's *d* [*d* 95% C.I.]	–	0.01 [-0.09 – 0.12]	–	-0.02 [-0.30 – 0.00]	–	0.02 [-0.00 – -0.04]	–	-0.03 [-0.05 – -0.00]
	n	594	935	1,340	189	31,284	30,460	52,847	8,897

[a]Reference groups for ease of interpretation across all comparisons are White and Under 40 years old. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Table 12

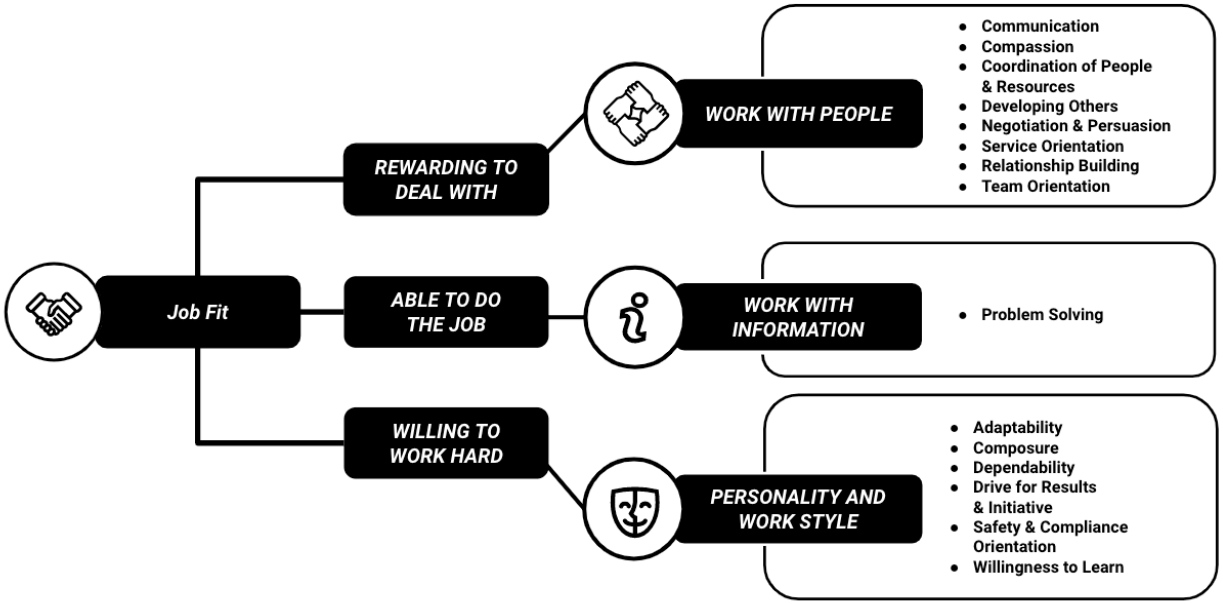
Samples 4 – 7 Subgroup Differences for Criterion Validation Samples with Applicant Demographics on Competency Model Scores

Study Name	Statistics	Gender		Ethnicity				
		Male[a]	Female	White[a]	Asian	Black	Hispanic	Two or more races
Sample 4	Cohen's *d* [*d* 95% C.I.]	–	-0.14 [-0.24 – -0.03]	–	–	0.10 [-0.03 – 0.24]	0.08 [-0.10 – 0.26]	–
	n	796	633	960	–	264	143	–
Sample 5	Cohen's *d* [*d* 95% C.I.]	–	-0.11 [-0.20 – -0.02]	–	–	-0.06 [-0.13 – 0.03]	-0.06 [-0.17 – 0.05]	–
	n	596	2,777	889	–	1,932	503	–
Sample 7	Cohen's *d* [*d* 95% C.I.]	–	0.06 [-0.01 – 0.13]	–	–	-0.20 [-0.27 – -0.13]	-0.14 [-0.31 – 0.03]	-0.02 [-0.22 – 0.19]
	n	2,960	1,080	1,406	–	2,271	145	8,567
Sample 8	Cohen's *d* [*d* 95% C.I.]	–	-0.09 [-0.15 – -0.04]	–	-0.09 [-0.26 – 0.07]	0.09 [0.05 – 0.14]	-0.01 [-0.07 – 0.04]	–
	n	1,446	10,672	2,940	148	6,030	2,034	–

[a]Reference groups for ease of interpretation across all comparisons are: Male and White. Positive Cohen's *d* values indicate the focal group has a higher mean than the reference group.

Figure 1

Competency Framework Guiding AVI-CA Development



Note. Rewarding to Deal With, Able to do the Job, and Willing to Work Hard framework components are adapted from Employability and Career Success: Bridging the Gap Between Theory and Reality (Hogan et al., 2013).

Appendix A

Competency Labels, Definitions, Sample Questions, Key Behaviors, and Theoretical Linkages with Other Competency Frameworks

Competency	Competency Definition	Competency Key Behaviors	Top O*NET Elements Content Linkages (Peterson et al., 1999)	Great 8 Competency Dimension Content Linkages (Bartram, 2005)
Adaptability	The ability to shift or change opinions, actions, or behaviors. Those ranking high in this competency can successfully adjust when faced with multiple demands, shifting priorities, rapid change, or ambiguity. Sample Question: *Tell me about a situation when you had to adapt to a substantial change you encountered while working on a project or school assignment. Please describe the situation, how you adapted to the change, and the outcome.*	• Sees the Positive in Change • Seeks to Understand Change • Adjusts Behavior to Accommodate Change • Drives the Change	• Adaptability/Flexibility	• 7.1.1 Adapting • 7.1.2 Accepting New Ideas • 7.1.5 Dealing with Ambiguity
Communication[a]	The ability to express ideas or a message in a clear and convincing manner. Those ranking high in this competency are able to listen attentively to ensure their message is understood and appropriately tailored to their audience.	• Delivers Clear & Concise Message • Uses Proper Grammar • Shares Information • Verifies Understanding • Engages Others • Tailors Message to Audience	• Oral Expression • Written Expression • Writing • Communicating with Supervisors, Peers, or Subordinates • Speaking	• 2.1.5 Listening • 2.1.7 Communicating Proactively • 3.3.1 Speaking Fluently • 3.3.2 Explaining Concepts and Opinions • 3.3.3 Articulating Key Points of an Argument • 3.3.4 Presenting and Public Speaking • 3.3.5 Projecting Credibility • 3.3.6 Responding to an Audience
Compassion	The ability to express genuine concern or interest in the well-being of others while showing empathy for another person's situation. Those ranking high in this competency consider the needs and difficulties people are facing and take an active interest in their feelings, challenges, and perspectives in order to assist or support them when possible. Sample Question: *Tell us about a time when you showed genuine concern for someone who was going through a difficult situation at work or school. Please describe the situation, your actions, and the outcome.*	• Acknowledges and Considers Others' Needs • Recognizes Own and Others' Emotions • Concerned for Others' Welfare • Sensitive to Individual Differences • Remains Courteous and Polite	• Concern for Others • Assisting and Caring for Others • Service Orientation • Coaching and Developing Others	• 2.1.8 Showing Tolerance and Consideration • 2.1.9 Showing Empathy • 2.1.10 Supporting Others • 2.1.11 Caring for Others

[a]Communication does not include an interview question since it was rated across competency-specific questions.

Appendix A con't

Competency Labels, Definitions, Sample Questions, Key Behaviors and Theoretical Linkages with Other Competency Frameworks

Competency	Competency Definition	Competency Key Behaviors	Top O*NET Elements Content Linkages (Peterson et al., 1999)	Great 8 Competency Dimension Content Linkages (Bartram, 2005)
Composure	The capacity to control emotions in the face of pressure, complaint, or failure while thinking clearly and logically despite the difficult situation. Those ranking high in this competency cope well with setbacks and remain calm when dealing with upset customers or co-workers. Sample Question: *Tell us about a time when you had to cope with a high-pressure or stressful work situation. Please describe the situation, your actions, and the outcome.*	• Effectively Manages Emotions • Recognizes Sources of Stress • Reacts Constructively • Recognizes Areas of Vulnerability • Maintains Order and Civility	• Self-control • Stress Tolerance • Resolving Conflicts and Negotiating with Others	• 7.2.1 Coping with Pressure • 7.2.2 Showing Emotional Self-control • 7.2.3 Balancing Work and Personal Life • 7.2.4 Maintaining a Positive Outlook • 7.2.5 Handling Criticism
Coordination of People & Resources	The ability to effectively utilize people, resources, and one's own time in order to accomplish an objective. Those ranking high in this competency work effectively with others to prioritize work activities and achieve optimal efficiency, both in long and short-term initiatives. Sample Question: *Tell us about a time when you were responsible for coordinating a team or a group. Please describe the situation, your actions, and the outcome.*	• Organizes Others to Achieve Goals • Understands the Broader Picture • Leverages Individuals' Unique Skill Sets • Distills Project to Manageable Tasks • Considers Input from Others	• Coordinating the Work and Activities of Others • Management of Personnel Resources • Management of Material Resources • Organizing, Planning, and Prioritizing Work • Time Management	• 1.2.1 Providing Direction and Coordinating Action • 1.2.4 Delegating
Dependability	The tendency to maintain high standards while following through on obligations by allocating sufficient time and focus to ensure high quality work. Those ranking high in this competency will exhibit integrity while taking pride in the quality and output of their work. They will carry an innate sense of honor and virtue and be able to monitor self-activities appropriately to achieve goals. Lastly, they will be willing to admit mistakes. Sample Question: *Tell us about when a new project challenged your ability to keep a prior commitment you made. Please describe the situation, your actions, and the outcome.*	• Sets a High Work Standard • Takes Principled & Ethical Approach • Takes Accountability & Feels Pride over Quality Work • Safeguards Resources	• Dependability • Integrity • Attention to Detail	• 6.1.4 Managing Resources • 6.3.4 Demonstrating Commitment • 2.2.1 Upholding Ethics and Values • 2.2.2 Acting with Integrity

Appendix A con't

Competency Labels, Definitions, Sample Questions, Key Behaviors and Theoretical Linkages with Other Competency Frameworks

Competency	Competency Definition	Competency Key Behaviors	Top O*NET Elements Content Linkages (Peterson et al., 1999)	Great 8 Competency Dimension Content Linkages (Bartram, 2005)
Developing Others	The ability to understand the strengths and developmental needs of others to help them reach their full professional potential. Those ranking high in this competency provide coaching, timely feedback, helpful counsel, and personal encouragement while challenging others with opportunities for professional development. Sample Question: *Tell us about as time you coached someone through performance challenges at work. Please describe the situation, your actions, and the outcome.*	• Identifies Others' Strengths and Addresses Skill Gaps • Holds Constructive Performance Management Conversations • Connects Performance Objectives to Organizational Goals • Provides Ongoing Coaching • Instructs on Performance Expectations	• Coaching and Developing Others • Training and Teaching Others • Instructing • Management of Personnel Resources • Leadership	• 1.2.7 Developing Staff • 1.2.3 Coaching • 1.2.5 Empowering Staff
Drive for Results / Initiative	The tendency to make continual strides toward exceeding goals or improving performance. Those ranking high in this competency focus on the bottom line and push themselves, as well as others, to achieve optimal results. Sample Question: *Tell us about a time you took initiative to improve an existing process at work. Please describe the situation, your actions, and the outcome.*	• Overcomes Obstacles • Establishes High Performance Goals • Aligns Goals • Focuses on Process Improvement • Monitors Performance of Self and Others	• Achievement/Effort • Initiative	• 6.2.2 Setting High Standards for Quality • 6.2.3 Monitoring and Maintaining Quality • 6.2.4 Working Systematically • 6.2.5 Maintaining Quality Processes • 6.2.6 Maintaining Productivity Levels • 6.2.7 Driving Projects to Results • 8.1.1 Achieving Objectives • 8.1.2 Working Energetically and Enthusiastically • 8.1.4 Demonstrating Ambition
Negotiation & Persuasion	The ability to work effectively with others to produce agreement on a course of action or outcomes that satisfy various interests. Those ranking high in this competency understand the motives, tactics, and goals of individuals while leveraging information to seek positive results. They are able to substantially influence the thoughts and actions of others to embrace a particular position, point of view, or course of action by establishing trust. Sample Question: *Sometimes a successful outcome requires compromise. Tell us about a time where you needed to negotiation with someone to achieve an objective. Please describe the situation, your actions, and the outcome.*	• Convinces Others • Adjusts Approach to Persuade • Provides Compelling Rationale • Neutralizes Tension and Gains Agreement	• Negotiation • Persuasion • Resolving Conflicts and Negotiating with Others	• 3.1.4 Managing Conflict • 3.2.2 Shaping Conversations • 3.2.3 Appealing to Emotions • 3.2.4 Promoting Ideas • 3.2.5 Negotiating • 3.2.6 Gaining Agreement

Appendix A con't

Competency Labels, Definitions, Sample Questions, Key Behaviors and Theoretical Linkages with Other Competency Frameworks

Competency	Competency Definition	Competency Key Behaviors	Top O*NET Elements Content Linkages (Peterson et al., 1999)	Great 8 Competency Dimension Content Linkages (Bartram, 2005)
Problem Solving	The ability to identify, analyze, determine causative factors, and find solutions for problems. Those ranking high in this competency are able to isolate the issue and use appropriate techniques to resolve the situation. Sample Question: *Tell us about a time you needed to make a decision quickly without all of the information you needed to solve a problem at work. Please describe the situation, your actions, and the outcome.*	• Recognizes and Addresses Issues • Anticipates Problems • Identifies Appropriate Solutions • Solves Problems	• Complex Problem Solving • Making Decisions and Solving Problems • Critical Thinking • Analytical Thinking • Deductive Reasoning	• 4.3.1 Analyzing and Evaluating Information • 4.3.2 Testing Assumptions and Investigating • 4.3.3 Producing Solutions • 4.3.4 Making Judgments • 4.3.5 Demonstrating Systems Thinking
Relationship Building	The ability to establish, build upon, enhance, and maintain friendly as well as mutually beneficial relationships. Those ranking high in this competency are able to use their established connections for professional development and effectiveness in the workplace. Sample Question: *Tell us about a time that it was important for you to build a relationship with a challenging coworker. Please describe the situation, your actions, and the outcome.*	• Seeks to Build and Nurture Relationships • Ensures Mutual Benefits • Aligns on Shared Goals • Establishes Personal Connection	• Establishing and Maintaining Interpersonal Relationships • Social Orientation • Developing and Building Teams • Cooperation	• 3.1.1 Building Rapport • 3.1.2 Networking • 3.1.3 Relating Across Levels
Safety & Compliance Orientation	The ability to adhere to rules regarding safety, compliance, regulations, and policies. Those ranking high in this competency recognize the rules that appropriately apply to their job role while being aware of potential risks in the workplace. They also encourage others to exhibit safe and appropriate working behaviors and raise concerns when there are unsafe conditions. Sample Question: *Tell us about a time when you witnessed an unsafe situation at work. Please describe the situation, your actions, and the outcome.*	• Follows Procedures • Inspects Work • Finds and Corrects Errors • Values Procedures • Intervenes on Issues	• Wear Common Protective or Safety Equipment such as Safety Shoes, Glasses, Gloves, Hearing Protection, Hard Hats, or Life Jackets • Wear Specialized Protective or Safety Equipment such as Breathing Apparatus, Safety Harness, Full Protection Suits, or Radiation Protection	• 6.3.5 Showing Awareness of Safety Issues • 6.3.1 Following Directions • 6.3.2 Following Procedures • 6.3.6 Complying with Legal Obligations
Service Orientation	The ability to interact effectively with customers, clients or other stakeholders to identify their needs, address their issues, meet expectations, and ensure satisfaction. Those ranking high in this competency go above and beyond to provide good service, and they handle difficult situations by empathizing and taking personal responsibility. Sample Question: *Tell us about a time when you provided service that exceeded a customer's expectations. Please describe the situation, your actions, and the outcome.*	• Builds Rapport • Determines Others' Needs • Resolves Problems • Delivers Timely and Effective Service • Ensures Others' Satisfaction	• Service Orientation • Concern for Others • Assisting and Caring for Others • Social Perceptiveness • Cooperation	• 6.2.1 Focusing on Customer Needs and Satisfaction

Appendix A con't

Competency Labels, Definitions, Sample Questions, Key Behaviors and Theoretical Linkages with Other Competency Frameworks

Competency	Competency Definition	Competency Key Behaviors	Top O*NET Elements Content Linkages (Peterson et al., 1999)	Great 8 Competency Dimension Content Linkages (Bartram, 2005)
Team Orientation	The ability to collaboratively define success in terms of the team or organization as a whole. Those ranking high in this competency focus on team accomplishment and recognition rather than personal gain while sharing successes equally to build a sense of togetherness. They design critical plans and delegate appropriately in order to help their team reach their goals. Sample Question: *Describe a time you joined a team and worked together to accomplish a common goal. Please describe the situation, your actions, and the outcome.*	• Prioritizes Team Success • Mitigates Team Conflict • Maximizes Team Effectiveness • Leverages Own and Team Members' Strengths • Shares Success	• Developing and Building Teams • Cooperation • Social Orientation • Establishing and Maintaining Interpersonal Relationships	• 2.1.2 Adapting to the Team • 2.1.3 Building Team Spirit • 2.1.4 Recognizing and Rewarding Contributions
Willingness to Learn	The motivation to absorb and apply new information, techniques, or procedures to work. Those ranking high in this competency are inquisitive about their job role and continually seek opportunities to close performance gaps or upgrade their individual skill level. They leverage appropriate resources, including their co-workers, to optimize performance and delight in sharing discoveries in best practices. Sample Question: *Tell us about a time when you decided to learn new skills outside of your responsibilities at work. Please describe the situation, your actions, and the outcome.*	• Open to Feedback • Closes Knowledge Gaps • Explores Opportunities for Growth • Applies New Knowledge • Promotes a Learning Environment	• Active Learning • Learning Strategies	• 5.1.1 Learning Quickly • 5.1.4 Encouraging and Supporting Organizational Learning • 8.1.3 Pursuing Self-development