ALGORITHMIC ABILITY PREDICTION IN VIDEO INTERVIEWS

by

Louis Hickman

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Psychological Sciences

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Sien Chieh Tay, Chair

Department of Psychological Sciences

Dr. Sang Eun Woo

Department of Psychological Sciences

Dr. Q. Chelsea Song

Department of Psychological Sciences

Dr. Michael A. Campion

Department of Management

Approved by:

Dr. Kimberly P. Kinzig

Dedicated to:

Ashley, Harper, and Conrad—for lifting me, supporting me, and making me capable of accomplishing more than I thought possible.

Mom and dad—for allowing me to dream.

Ollie and Eleanor—for helping me know my worth.

ACKNOWLEDGMENTS

I am grateful to Dr. Tay and the members of my committee for their guidance and feedback throughout this process. This would not have been possible without the action of Dr. Sang Eun Woo. Thank you to the many undergraduate students who helped with data collection and provided interviewer ratings. Thank you to Dr. Sidney D'Mello, Brandon Booth, and Krishna Subburaj for their help generating the transcripts, extracting paraverbal behaviors, and extracting nonverbal behaviors. This work was supported by a National Science Foundation Early Concept Grant for Exploratory Research (grant number 1921111) and a Society for Industrial and Organizational Psychology Anti-Racism Grant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Automated video interviews (AVIs) use machine learning algorithms to predict interviewee personality traits and social skills, and they are increasingly being used in industry. The present study examines the possibility of expanding the scope and utility of these approaches by developing and testing AVIs that score ability from interviewee verbal, paraverbal, and nonverbal behavior in video interviews. To advance our understanding of whether AVI ability assessments are useful, I develop AVIs that predict ability (GMA, verbal ability, and interviewer-rated intellect) and investigate their reliability (i.e., inter-algorithm reliability, internal consistency across interview questions, and test-retest reliability). Then, I investigate the convergent and discriminant-related validity evidence as well as potential ethnic and gender bias of such predictions. Finally, based on the Brunswik lens model, I compare how ability test scores, AVI ability assessments, and interviewer ratings of ability relate to interviewee behavior. By exploring how ability relates to behavior and how ability ratings from both AVIs and interviewers relate to behavior, the study advances our understanding of how ability affects interview performance and the cues that interviewers use to judge ability.

INTRODUCTION

A revolution is occurring in pre-employment assessment. The emergence of big data approaches, including machine learning (ML) algorithms, has opened up new ways of screening job candidates and assessing their knowledge, skills, abilities, and other characteristics (KSAOs; Chamorro-Premuzic et al., 2017). Specifically, vendors are marketing, and organizations are adopting tools that use games, virtual reality, and computer-extracted interviewee behaviors to assess KSAOs automatically (Raghavan et al., 2019). Collectively, these tools tend to use applicant in situ behaviors as inputs to supervised ML algorithms to score KSAOs. These new approaches hold promise for advancing pre-employment assessment science and practice in two primary ways. First, by increasing the efficiency and decreasing the cost of assessment and selection. Second, by providing new methods for advancing our conceptual understanding of assessee performance and how it translates into assessor ratings.

One emerging assessment garnering significant interest is automated video interviews (AVIs)—for example, one vendor had already conducted over a million AVIs two years ago (Harwell, 2019), and over a half-dozen vendors offer similar products (Raghavan et al., 2019). AVIs use interviewee verbal, paraverbal, and nonverbal behaviors as inputs to ML algorithms to score interviewee KSAOs (Hickman, Bosch, et al., 2021). Organizations have begun adopting AVIs due to several potential benefits, including that: applicants may react more positively to interviews than personality and ability tests (which reduces the likelihood of litigation; Hausknecht et al., 2004); AVIs are highly scalable; AVIs can provide considerable time and cost savings compared to manual screening methods; and vendors claim that AVIs improve the quality of new hires and reduce time to hire (Oswald et al., 2020). Unfortunately, research on AVI benefits is lacking, resulting in a science-practice gap where psychology and management research lag behind application (Ones et al., 2017; Rotolo et al., 2018).

Initial research has emerged to suggest the viability of using AVIs to assess interviewee KSAOs (e.g., Nguyen et al., 2014). However, the existing research is limited because it has tended to focus on suboptimal predictors of job performance such as five factor model (FFM) and HEXACO personality traits (e.g., Hickman, Bosch, et al., 2021; Jayaratne & Jayatilleke, 2020). Personality traits are only relevant to performance in specific jobs (e.g., Judge & Zapata, 2015),

and the potential utility of AVIs can be increased by focusing on one of the strongest predictors of performance across occupations—ability.

Ability, or often general mental ability (GMA or *g*), has long been recognized as the best predictor of both job and training performance across occupations (Schmidt & Hunter, 1998) but especially so in complex jobs (Hunter et al., 2006). Indeed, GMA is a better predictor of task and overall job performance than the combined FFM traits (Gonzalez-Mulé et al., 2014), and interviewers frequently judge interviewee GMA, explicitly or implicitly (Huffcutt et al., 2001). Conceptually, verbal ability is the specific ability that is most relevant to interview performance, is a key component of GMA as it is involved in a variety of specific abilities (Carroll, 1993), and is a strong predictor of both job *and* interview performance (König et al., 2007; Lang et al., 2010; Melchers et al., 2009).

The lack of focus on ability in AVIs is reflected more broadly in employment interview research. For example, many studies have explored how personality traits relate to interview performance and how interview performance leads to personality judgments by interviewers (e.g., Bourdage et al., 2018; Peeters & Lievens, 2006; Van Iddekinge et al., 2007). Yet, I could find no studies that examined the effects of ability on interview performance, nor any investigations of how interview performance relates to ability judgments. This gap exists even though interview ratings of *other constructs* correlate highly with ability (Roth & Huffcutt, 2013), suggesting that ability plays a pervasive, yet largely unexplored, role in interview performance.

Therefore, the present study aims to accomplish two primary goals. First, to investigate the psychometric properties of AVI ability assessments. Second, to advance our understanding of how ability relates to interview performance, and how interview performance relates to interviewer ratings of ability. To accomplish this, I train ML models to predict GMA, verbal ability, and interviewer-rated intellect on a collection of mock interviews and examine their psychometric properties. In doing so, the present study makes four primary contributions to the personnel assessment and selection literature.

First, the study examines the reliability of AVI ability assessments. The reliability of AVI assessments has largely been ignored in prior research (for an exception, see Hickman, Bosch, et al., 2021). Several forms of reliability will be investigated, including: 1) inter-algorithm reliability, akin to interrater reliability (Sajjadiani et al., 2019), which isolates variance specific to the mathematical model used by correlating the predictions from different ML algorithms to assess

their consistency; 2) split-half reliability, wherein variance specific to the interview questions asked will be isolated by making predictions separately on the odd and even numbered questions, then correlating those predictions; and 3) test-retest reliability, which isolates variance specific to occasions by correlating ability predictions made for a sample of participants who participated twice in the mock video interview.

Second, I examine the extent to which algorithmic ability predictions converge with ability multiple-choice test scores, interviewer-rated intellect, and commonly used proxies for ability (i.e., self-reported standardized test scores and academic performance). Most prior research into AVIs has only investigated convergence with the measure the ML model was trained to predict (what I call *internal* convergence), yet examining convergence with other, similar measures (what I call *external* convergence) provides more substantial evidence that can support the validity of algorithmic predictions in personnel selection (AERA et al., 2014; SIOP, 2018). I also compare how ability predictions and observed ability (i.e., test scores and interviewer ratings) converge with proxies for ability to advance our understanding of how construct validity is affected by replacing traditional measures with ML models. Further, I investigate ethnic and gender group differences and investigate the extent of measurement bias in the AVI ability scores.

Third, the study will investigate the ability of AVI predictions to discriminate between ability and personality traits. Measurement discrimination represents another piece of evidence to support the proposed use of such assessments. Yet, existing work using ML to predict personality traits has rarely provided such evidence (Bleidorn & Hopwood, 2019; Tay et al., 2020), and only one study of AVIs has done so (Hickman, Bosch, et al., 2021). I will train AVI personality and hireability assessments to predict interviewer ratings of these constructs. The intercorrelations among predicted and observed values for ability and personality will be evaluated using a multitrait-multimethod (MTMM) matrix. Evidence of stronger convergent than discriminant correlations, as well as expected interrelationships among personality traits and ability, would further support the validity of algorithmic ability predictions.

Fourth, the study will draw on the Brunswik lens model (1956) and the lens model equations (Hursch et al., 1964; Tucker, 1964) to examine how ability manifests behaviorally in an interview, the behavioral cues related to AVI ability assessments, and the behavioral cues related to interviewer intellect judgments. Interviews are commonly used to assess ability (Huffcutt et al., 2001). Yet, the accuracy of interviewer ability judgments (in terms of convergence with ability

test scores) has rarely been investigated, and the validity of behavioral cues used to judge ability has, to my knowledge, not been examined. Further, the lens model equations enable a decomposition of convergent correlations to advance our understanding of interpersonal ability perceptions. Specifically, the equations enable the examination of which raters use cues more consistently and validly beyond merely examining convergence with ability test scores. If, compared to interviewers, algorithms show greater convergence with tested ability, use cues more consistently, and use more valid cues to predict ability, algorithms would represent an empirically and conceptually attractive alternative to interviewer judgments of ability.

I begin by introducing ability and review its connection to interviewee performance and behavior more broadly. Next, I critically review past research on AVI KSAO predictions. Then, I draw on psychometrics research to identify methods for advancing our understanding of the reliability and validity of AVI ability assessments that expand upon prior AVI research. Further, I review proximal behaviors related to ability and explain how the Brunswik lens model and the lens model equations can be used to advance our understanding of the validity of ability ratings.

Ability

Ability is strongly predictive of outcomes in life and at work, particularly in more complex settings (Gottfredson, 1997; Lubinski, 2004). General mental ability represents the broadest operationalization of ability. It was first conceptualized 12 decades ago (Spearman, 1904), can be described as "a highly general information-processing capacity that facilitates reasoning, problem solving, decision making, and other higher order thinking skills" (p. 81) and is measured as the variance shared by a variety of cognitive tests (Gottfredson, 1997). Factor analyses of specific ability tests reveal that approximately half of the variance in specific abilities is accounted for by this general factor (Wechsler, 1997).

However, recent research has explored whether it is necessary to combine multiple tests to score GMA, or if specific abilities can be validly used on their own for personnel selection. Verbal ability, one of Thurstone's (1938) primary human abilities, appears to have higher relative importance for predicting job performance than GMA (Lang et al., 2010) and is as important as GMA for predicting occupational prestige (Lang & Kell, 2019). Several well-known tests of GMA tend to oversample from verbal ability (e.g., Wechsler Adult Intelligence Scale; Wonderlic Personnel Test), suggesting that the well-established relationship between GMA and workplace

outcomes may primarily be a function of verbal ability (Schneider & Newman, 2015; Lang & Kell, 2019). Further, Carroll (1993) suggested that verbal ability is the most important specific ability because many other ability factors require verbal ability, and virtually all tests presuppose the test taker's knowledge of their native language. Therefore, although most research has focused on the effects of GMA on workplace outcomes, the present study also investigates whether verbal ability can be algorithmically inferred in video interviews.

Ability is also often judged, implicitly or explicitly, by interviewers (Huffcutt et al., 2001). Employers may assess ability with interviews instead of tests for several reasons, including: logistic concerns (i.e., time and cost required for proctored tests); legal concerns (i.e., adverse impact); because applicants react more positively to interviews than ability tests; and because hiring manager's decisions are influenced more by interviewer-rated ability than ability test scores (Hausknecht et al., 2004; Huffcutt et al., 2001; Lievens et al., 2005). However, interviewer-rated ability is a much less valid predictor of job performance than ability test scores (Huffcutt et al., 2001).

So, although ability test scores are widely considered the best operationalization of ability (Chamorro-Premuzic & Furnham, 2005), the present study uses GMA and ability test scores as well as interviewer ratings of intellect. These three operationalizations of ability are then used as the "ground truth" (or *y* variable) to train ML models.

Benefits of Algorithmic Ability Prediction in Video Interviews

As mentioned in the introduction, AVIs have focused primarily on interviewer-rated FFM traits, yet the utility of AVIs could be improved by additionally assessing ability. Conceptually, ability represents what people *can* do, or their maximal performance, while personality traits represent what people *will* tend to do, or their typical performance (Cronbach, 1990). Therefore, ability and personality traits can be used complementarily for predicting workplace behavior.

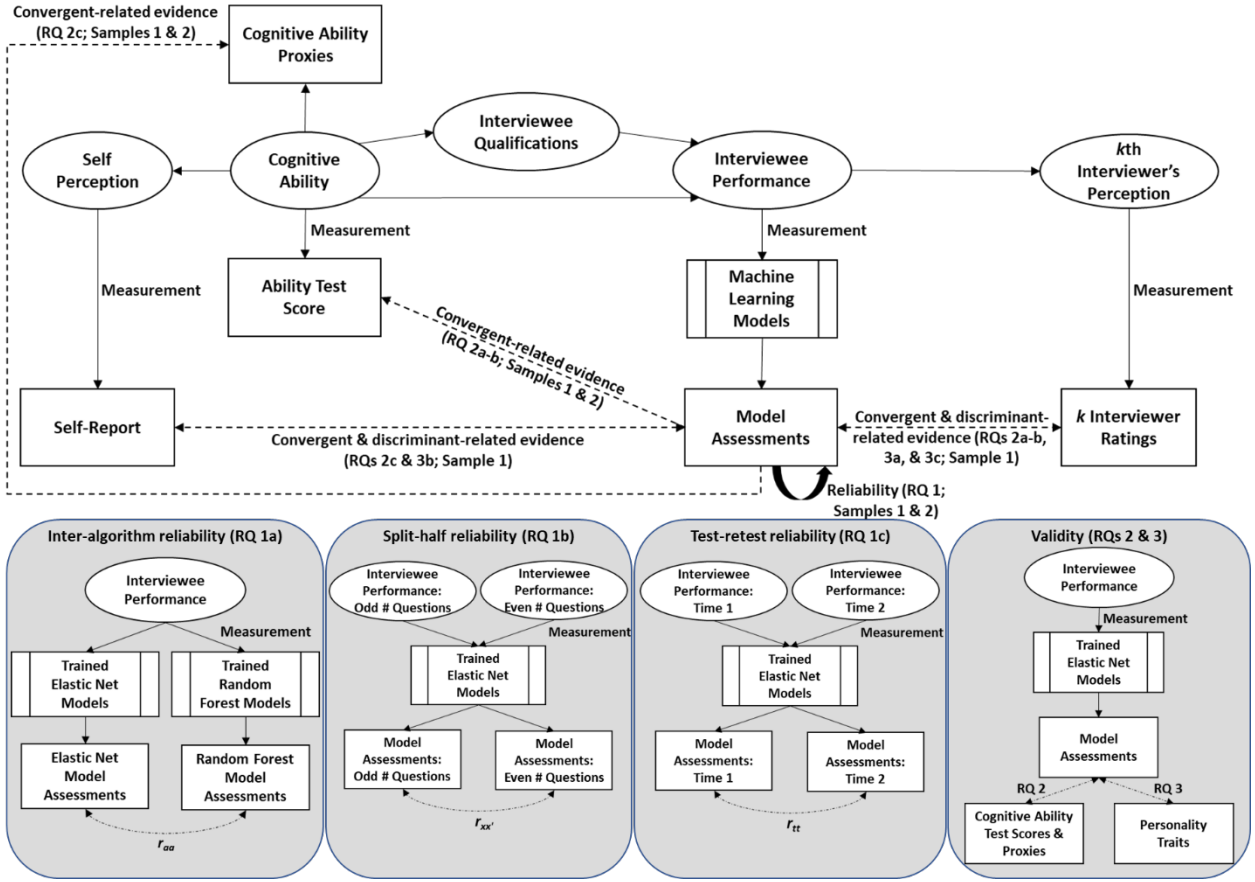
Ability is considered the strongest predictor of both job and training performance (Schmidt & Hunter, 1998). This is especially true in complex jobs, where ability accounts for over half of the variance in job performance after correcting for range restriction (Hunter et al., 2006). Ability is also the strongest predictor of training performance (Schmidt & Hunter, 1998), which is unsurprising because ability tests originated in efforts to differentiate competent from incompetent students (Binet & Simon, 1905). The positive effect ability has on training performance indirectly

affects job performance because higher ability improves one's ability to gain declarative and procedural job knowledge (Campbell et al., 1993). As a result, ability also has a strong effect on extrinsic career success (i.e., income and occupational prestige; Judge et al., 1999; Ng et al., 2005).

One potential reason that ability, as operationalized via tests, is such a strong predictor of important outcomes is that ability tests are less subject to individual biases compared to self-reports and interviews. Self-reports are contaminated by self-enhancing biases and self-presentation effects (Vazire, 2010), and attempts to measure and correct for these biases are generally unsuccessful (e.g., Piedmont et al., 2000). Meta-analytic estimates suggest that self-reported ability shares only 10.89% of the variance with ability test scores in low-stakes settings (Freund & Kasten, 2012). In fact, at times, observer reports of ability based on thin slices of behavior can share more variance with ability test scores than self-reports (e.g., Borkenau & Liebler, 1993). However, as mentioned, interviewer ratings of ability are less valid predictors of job performance than test scores (Huffcutt et al., 2001).

Ability tests have been developed for more than a century and require individuals to demonstrate the ability they possess, making them a more objective operationalization (Chamorro-Premuzic & Furnham, 2005) of ability that is not prone to interviewer biases or information processing limitations and is less fakable than self-reports. Together, this evidence suggests the value of using ability test scores as the ground truth for ML algorithms because they are 1) the best-established predictors of training and job performance and 2) measured in an objective way. Additionally, as illustrated in Figure 1, ability test scores are more directly related to ability than either self-reports or interviewer-ratings.

As Figure 1 also shows, interviewer-rated ability is more directly related to interviewee behavior than are ability test scores. The relationship between ability test scores and interview performance is mediated by the latent, unobservable ability construct and partially through interviewee qualifications, whereas interviewer perceptions of interviewee KSAOs are based directly on interview performance. Therefore, interviewer-ratings are likely to be more accurately modeled than ability test scores, similar to how interviewer-rated personality traits are more accurately modeled than self-reported traits (Hickman, Bosch, et al., 2021). What is unknown, however, is whether AVI models of ability test scores or interviewer ratings will capture more ability relevant variance.



Note. Adapted from Hickman et al. (2021).

Figure 1. Construct validation framework for algorithmic ability assessment.

Links Between Ability and Outcomes

Ability in interviews. The model of interviewee performance (Huffcutt et al., 2011) posits that ability affects interviewee performance in two ways (as shown in Figure 1): directly and indirectly through the interviewee's job relevant qualifications (i.e., declarative knowledge, procedural knowledge, and motivation). The same way that job performance is best conceptualized as behaviors (not outcomes of those behaviors; Campbell et al., 1993), interviewee performance consists of verbal behaviors (i.e., the answers given to interview questions), paraverbal behaviors (i.e., the pitch, speech rate, and voice quality associated with the verbal behavior), and nonverbal behaviors (i.e., the gestures and facial expressions exhibited in the interview). Table 1 provides examples of how verbal, paraverbal, and nonverbal behaviors are operationalized in the present study.

As the model of interviewee performance (Huffcutt et al., 2011) suggests, ability has large effects on interviewee performance. Employment interviews are cognitively demanding, as they require interviewees to simultaneously respond to novel questions, recall past experiences, and engage in impression management (König et al., 2007). Employment interview ratings exhibit a corrected meta-analytic correlation of .42 with GMA test scores, making GMA nearly as strong a predictor of interview performance as job performance (Roth & Huffcutt, 2013) and suggesting that interviewee behavior is strongly affected by ability. Although GMA's indirect effect on interviewer ratings has received research attention, the direct effect of GMA on interviewee performance has been understudied. For instance, although research has explored how personality traits affect impression management tactics (e.g., Bourdage et al., 2018; Peeters & Lievens, 2006; Van Iddekinge et al., 2007), similar research has not investigated how GMA affects impression management tactics.

Some research has demonstrated that ability improves interview performance through the interviewee's ability to identify criteria (ATIC). ATIC is an individual difference that determines how accurately one can identify the behaviors required for success in evaluative situations (Speer et al., 2014). Interviews are highly ambiguous for interviewees because they are not always informed about the criteria used to evaluate their performance. Interviewees who successfully identify the criteria can adjust their behavior accordingly and receive higher ratings for doing so, both in interviews and assessment centers (Ingold et al., 2015; König et al., 2007; Melchers et al.,

Table 1. Example Operationalizations of Verbal, Paraverbal, and Nonverbal Behaviors

Verbal Behaviors	Paraverbal Behaviors	Nonverbal Behaviors

Descriptive
• Word count
• Unique word count (lexical diversity)
• Mean number of syllables per word
• Readability
Closed Vocabulary Text Mining
• Use LIWC to count conceptually related words to measure constructs. For example:
  • Cognitive processes
  • Causation
  • Certainty
  • Achievement words
  • Present focus words
  • Perceptual processes
  • Work
  • Money
	Pitch
	• Amplitude
	• Volume
	• Bandwidth
	• Frequency
	• Speech rate
	• Duration of Pauses
	• Voice quality/smoothness
	• Stop words per second
	• Filler words per second
	Note: Each is described by mean and standard deviation
		• Facial expressions
		  • Facial action unit intensity (19)
		  • Blinks
		  • Anger
		  • Contemp
		  • Disgust
		  • Joy
		  • Fear
		  • Sadness
		  • Surprise
		  • Positivity
		  • Negativity
		• Body languate
		  • Eye contact
		  • Head pose
		  • Head orientation

Table 1 continues

Verbal Behaviors	Paraverbal Behaviors	Nonverbal Behaviors
• Analytical • Authentic • Clout • Tone • Used mean and standard deviation across all questions for each LIWC category Open Vocabulary Text Mining • Bag of words • 1-3 word phrases (1,000s) • 50 Topics		Note: Each is described by mean and standard deviation

Note. LIWC = Linguistic Inquiry and Word Count.

2009; Speer et al., 2014). Indeed, researchers have found that ATIC mediates ability's effects on interview and assessment center performance (Kleinmann et al., 2011). Interestingly, verbal ability, but not matrix completion scores (usually considered a measure of either GMA or fluid intelligence; Carroll, 1993), predicts ATIC, and verbal ability and ATIC predict assessment center and interview scores at similar magnitudes (Griffin, 2014; König et al., 2007; Melchers et al., 2009). Given this, interviewees with high ability, and particularly verbal ability, are likely better able to guess and enact behaviorally appropriate responses, although how this manifests in proximal behaviors has not yet been studied.

Distal outcomes of ability. Although the relationship between ability and interviewee behavior has rarely been examined, research has investigated how ability affects behavior in other situations. Most research has linked ability to distal outcomes, including the career and workplace outcomes previously mentioned, but also organizational citizenship behaviors (Gonzalez-Mule et al., 2014), life satisfaction (Gonzalez-Mulé et al., 2017), physical health (Judge et al., 2010), and longevity (Gottfredson & Deary, 2004). If ability affects so many distal outcomes, it appears likely that ability causes differences in proximal behaviors that mediate the distal effects of ability. I next introduce AVIs before discussing the Brunswik lens and then review research on the proximal behaviors caused by ability.

Automated Video Interviews

Primarily computer scientists have begun investigating the potential of using ML algorithms to infer interviewee personality traits, social skills, and hireability in video interviews. These studies use a combination of verbal (i.e., what people say), paraverbal (i.e., how they say it, such as voice quality and speech rate), and/or nonverbal (i.e., facial expressions and gestures) behaviors to predict interviewer-rated attributes. Whereas studies using digital footprints to infer FFM traits have generally utilized self-reports as ground truth (for a meta-analysis, see: Azucar et al., 2018), the studies of AVIs have relied almost exclusively on interviewer ratings as ground truth and are summarized in Table 2 (note that Jayaratne & Jayatilleke, 2020, which used self-reports, is a study of automated text-based, not video, interviews; Hickman, Bosch, et al., 2021 used both self-reports and interviewer-ratings). Nearly all of these studies come from other fields (for an exception, see Hickman, Bosch, et al., 2021), so there is much to learn from their

computational methods, although some other methodological aspects of their studies fall short of the best practices in applied psychology.

Several of these studies have demonstrated an impressive ability to replicate interviewer ratings of interviewee KSAOs. For instance, several studies have achieved cross-validated convergence with overall assessments and hiring recommendations of $r \sim .60$ (e.g., Naim et al., 2018) and with personality traits including conscientiousness and extraversion rs > .5 and .6, respectively (e.g., Hickman, Bosch, et al., 2021). The winning team of the ChaLearn First Impressions challenge achieved high convergence with all personality ratings, ranging from $r = .58$ for agreeableness to $r = .73$ for conscientiousness (Ponce-López et al., 2016). However, except for Hickman et al. (2021), no reliability evidence has been provided, and internal convergence (i.e., convergence with the ground truth), is the only validity evidence provided.

Several additional shortcomings of these studies deserve note. First, several studies have used in-person interviews with an interviewer present (e.g., Muralidhar et al., 2016; Naim et al., 2018; Nguyen et al., 2014). These studies lack ecological validity, as algorithmic assessments are used in the context of one-way (asynchronous) video interviews. Additionally, one study used *interviewer* behaviors as predictors of interviewee attributes (e.g., Nguyen et al., 2014), thereby contaminating the operationalization of interview performance with irrelevant variance (i.e., construct contamination; SIOP, 2018). Relatedly, several studies have used only one or two types of behaviors, sometimes entirely ignoring what is said (i.e., verbal behavior; Biel et al., 2013; Muralidhar et al., 2016; Nguyen & Gatica-Perez, 2016; Nguyen et al., 2014; Rasipuram & Jayagopi; 2019). The model of interviewee performance (Huffcutt et al., 2011) states that interview performance, or the behaviors that interviewees exhibit, is comprised of verbal (i.e., their answers to interview questions), paraverbal (i.e., how they sound when answering), and nonverbal (i.e., their posture, gestures, and facial expressions) behaviors. Therefore, any operationalization of interviewee performance that does not include all three types of behavior may be considered deficient (i.e., construct deficiency; SIOP, 2018). Notably, however, if AVIs focus on only one type of behavior, verbal behavior is likely to be the most acceptable to applicants. Interview best practices suggest using behaviorally anchored rating scales (e.g., Campion et al., 1997) to improve validity and fairness. Behaviorally anchored interview ratings are supposed to be based entirely on what interviewees say—not how they say it or their nonverbal behaviors.

Table 2. Summary of Past Automated Interviews Research

Authors (Year)	N	Interview Setting	Interview Characteristics	Types of Predictors	Constructs Assessed	CV Strategy	Best CV Accuracy
Biel et al. (2013)	408	Video blogs	None provided	V, NV, & PV	E, A, C, N, O	10-fold with inner 5-Fold CV	R^2: E=.48; A=.39; C=.22; ES=.23; O=.17
Chen et al. (2016)	36	Proctored, one-way	12 PBQs	V, NV, & PV	BARS scores, E, A, C, N, O, & Hiring Rec.	LOOCV with inner 10-fold CV	*r*s: BARS=.43; E=.44; A=.38; C=.34; ES=.40; O=35; Hiring=.41
Chen et al. (2017)	260	Remote, one-way	8 PBQs	V, NV, & PV	E, A, C, N, O, & Hiring rec.	80/20 train/test split with inner 5-fold CV	F1 scores: E=.78; A=.84; C=.86; N=.83; O=.81; Hiring=.66
Hickman et al. (2021)	1,073	Remote, one-way	Sample 1: 1 unstructured Q; Sample 2: 1 unstructured Q and 2 PBQs; Sample 3: 5 PBQs	V, NV, & PV	E, A, C, N, O	Nested 10-fold cross-validation and cross-sample cross-validation	*r*s: E=.65; A=.44; C=.52; N=.32; O=.41
Jayaratne & Jayatilleke (2020)	12,183	Text-based	5-7 PBQs	V	HEXACO traits	80/20 train/test split (N is of the test group)	*r*s: Honesty-humility=.44; eXtraversion=.34; Emotionality=.33; A=.28; C=.44; O=.50
Muralidhar et al. (2016)	169	In-person, two-way	7 questions, 3 unstructured	NV & PV	Overall, motivated, competent, hard-working, sociable, enthusiastic, positive, communicative, con-cise, persuasive	LOOCV with inner 10-fold CV	R^2: Overall=.32; motivated=.29; competent=.18; hardworking=.15; sociable=.19; enthusiastic=.34; positive=.30; communication=.25; concise=.14; persuasive=.20

Table 2 continues

Authors (Year)	N	Interview Setting	Interview Characteristics	Types of Predictors	Constructs Assessed	CV Strategy	Best CV Accuracy
Naim et al. (2018)	69 (2 per participant for 138)	In-person, two-way	5 questions, 3 unstructured and 2 PBQs	V, NV, & PV	Overall, structured no fillers, pauses, focused, not awkward, speech rate authentic, calm, not stressed, eye contact, excited, engaged, friendly, smiled, Hiring rec.	1000 trials of 80/20 train/test split	*r*s: Overall=.62; structured=.64; no fillers=.59; pauses=.58; focused=.58; no awkward=.52; speech rate=.46; not stressed=.26; eye contact=.33; excited=.79; engaged=.75; friendly=.73; smiled=.71; Hiring=.65
Nguyen & Gatica-Perez (2016)	939	Video resumes	123.5 sec median length	NV & PV	Overall, E, A, C, N, O, Social skills	10-fold CV	R^2: Overall=.18; E=.27; A=.06; C=.03; N=.00; O=.20; Social skills=.21
Nguyen et al. (2014)	62	In-person, two-way	8 questions, 4 unstructured	NV & PV	Communication Skills, Persuasion Skills, C, Stress Resistance, & Hiring rec.	LOOCV	R^2: Communication=.02; Persuasion=.12; C=.04; Stress resistance=.27; Hiring=.36
Rasipuram & Jayagopi (2018)	100 (2 per participant)	1) Remote, one-way; 2) In-person, two-way, Remote, one-way	5 PBQs from a pool of 100	V, NV, & PV	Communication skills	Leave 5 out CV	Accuracy (for classifying low performers): .82 remote; .86 in-person

Table 2 continues

Authors (Year)	N	Interview Setting	Interview Characteristics	Types of Predictors	Constructs Assessed	CV Strategy	Best CV Accuracy
Rasipuram & Jayagopi (2019)	251	Remote, one-way	5 PBQs from a pool of 100	NV, & PV	Communication skills	45 held out for testing	*r*: Communication=.11

Note. PBQ=past behavioral question. V=verbal behavior. PV=paraverbal behavior. NV=nonverbal behavior. E=extraversion. A=agreeableness. C=conscientiousness. N=neuroticism. O=openness to experience. Rec.=recommendation. CV=cross-validation. LOOCV=leave one out cross-validation. Jayaratne & Jayatilleke (2020) trained their models on self-reports.

Second, many of these studies have relied on small samples, with as few as 36 to 69 participants (Chen et al., 2016; Naim et al., 2018; Nguyen et al., 2014). Relatedly, although the ChaLearn First Impressions personality prediction competition included many videos, it relied on short, 15-second clips pulled from YouTube suitable only for judging first impressions (Ponce-Lopez et al., 2016). Yet, this project was referred to as a job candidate screening challenge (Liem et al., 2018). AVI models developed on small samples are unlikely to generalize due to sampling error and sample homogeneity (Bleidorn & Hopwood, 2019). AVI models developed on short clips are modeling strangers' first impressions, which are the least accurate form of observer ratings (Connelly & Ones, 2010) and, if deployed to assess actual job applicants, would likely be perceived as unjust (Yankov et al., 2020).

Third, many of these studies have used ad-hoc, single-item scales (Muralidhar et al., 2016; Naim et al., 2018; Nguyen & Gatica-Perez, 2016; Nguyen et al., 2014; Rasipuram & Jayagopi, 2018) and/or raters who have not undergone frame of reference training (Biel et al., 2013; Ponce-Lopez et al., 2016; Naim et al., 2018; Nguyen & Gatica-Perez, 2016; Rasipuram & Jayagopi, 2018). The use of ad-hoc, single-item scales calls into question the reliability and validity of the ground truth, as the scales have not been validated, nor is it always clear what they measure. Similarly, having interviewers undergo frame of reference training is essential because it increases interrater reliability and the validity of ratings (Campion et al., 1997). Further, none of the studies except Hickman, Bosch, et al. (2021) have provided evidence of the validity of the interviewer ratings, such as by demonstrating that they converge with self-reports, test scores, academic outcomes, or workplace criteria. Together, these shortcomings warrant expanded investigations of AVIs.

To my knowledge, at least two datasets have been used to train and test ML ability models. Kosinski et al. (2013) worked with the MyPersonality dataset and used Facebook likes to predict Raven's Standard Progressive Matrices scores. Their 10-fold cross-validated accuracy $r = .39$ ($N = 1{,}350$), which was slightly less accurate than their predictions of self-reported extraversion and openness, yet more accurate than their predictions of agreeableness, conscientiousness, or emotional stability.

Sergienko and Schmitt (2015), Fernandez-Martinez et al. (2012), and Zablotskaya (2015) used the same data to predict scores on a subset of the Wechsler Intelligence Test (1997). The datasets were generated by having participants watch a short film, then having the participants either a) explain what they saw as if speaking to a friend (monologue) or b) discuss the film with

another person (dialogue). I focus on Sergienko and Schmitt's (2015) investigations of using text mining to classify participants as either high or low intelligence in the two datasets. In the monologue dataset, baseline accuracy was 60% because 60 of the 100 participants were classified as high intelligence, and baseline accuracy was 59% in the dialogue dataset because 54 of 91 participants were classified as high intelligence. Using leave-one-out cross-validation, the highest accuracy they obtained was .67 for the monologues and .71 for the dialogues, each only somewhat higher than the baseline accuracy.

Understanding the Reliability and Validity of Algorithmic Ability Assessment

Using behavioral cues to predict KSAOs with ML algorithms represents a form of empirical criterion keying. Empirical criterion keying involves using a known outcome, the *ground truth*, such as group membership (e.g., leaders and non-leaders; Ozer & Reise, 1994) to select a set of scale items that best predict that outcome (e.g., the Minnesota Multiphasic Personality Inventory; Hathaway & McKinley, 1943). In algorithmic KSAO predictions, predictor weights are keyed using an existing measure of the target construct (Bleidorn & Hopwood, 2019). However, the sole focus on convergence with the ground truth in item selection and weighting is an atheoretical process that can cause the predictions to have poor internal consistency reliability (Loevinger, 1957; Simms, 2008). This occurs because items (in the present study, behavioral cues) are selected and weighted based only on their ability to predict the ground truth, resulting in high item heterogeneity (Simms, 2008). Examining the reliability of algorithmic ability inferences is important because reliability is traditionally thought to set an upper limit on validity (Cronbach, 1990).

Reliability. Although Cronbach's alpha is commonly used to examine scale reliability by correlating together scale items, different approaches must be used to examine the reliability of algorithmic ability assessments. Specifically, inter-algorithm reliability, split-half on interview behavior, and test-retest reliability are relevant indices. Inter-algorithm reliability (r_{aa}) is a form of interrater reliability (Sajjadiani et al., 2018) where multiple algorithms are treated as 'raters' and their predictions are correlated to assess their level of consistency. Low inter-algorithm reliability

indicates that the relationships between predictors and the ground truth are relatively weak. In the present study, models were developed using elastic net regression and random forest (prior AVI work has used both—e.g., Chen et al., 2017; Hickman, Saef, et al., 2021), then their predictions were correlated to examine their consistency. Low correlations between predictions from separate algorithms suggests that algorithm-specific variance is contaminating the assessment (Tay et al., 2020).

Although the items selected during empirical keying will tend to have low intercorrelations, the predictions across subsamples of behavior should be positively correlated. A form of split-half reliability based on making two assessments from a larger sample of behavior, rather than from two sets of scale items, can investigate this. One study that used Facebook posts to predict self-reported personality (Park et al., 2015) presented evidence of automatic KSAO assessment internal consistency in the form of split-half reliability. They split individuals' Facebook feeds into multiple segments of at least 1,000 words and made personality predictions for each segment. They found intercorrelations ranging from .61 to .71 across these assessments. In the present study, I calculated behavioral cues for the odd- and even-numbered questions of an interview separately, generated algorithmic predictions of ability, then correlated them to calculate split-half reliability, r_{xx}'. This speaks to the ML models' generalizability across the universe of interview questions (cf. Hickman, Bosch, et al., 2021) while simultaneously constraining the amount of behavior made available to them. Split-half reliability may partially be a function of the type of behaviors used by the algorithm because test-retest reliability tends to be lower for AVI models that use *n*-grams and higher for models that use conceptual categories (which tend to vary less across situations) like those calculated by LIWC (Hickman, Bosch, et al., 2021). Such effects are likely similar whether analyzed across time or across interview questions.

Test-retest reliability (r_{tt}) is fundamental to testing, as the primary concern of reliability is whether a person's scores would converge if tested twice (Cronbach, 1990). However, test-retest reliability is largely an unknown characteristic of employment interviews, with the only study I am aware of finding that behavioral, situational, and experience/interest interviews exhibited test-retest reliability r = .30, .35, and .26, respectively, over a one-year interval (Schleicher et al., 2010). The only evidence of AVI reliability so far was provided by Hickman et al. (2021), who found that test-retest reliability could be high when AVIs were trained to score self-reported agreeableness (r_{tt} max = .85) or emotional stability (r_{tt} max = .85), as well as for interviewer-rated extraversion

(r_{tt} max = .74), conscientiousness (r_{tt} max = .76), and to a lesser extent, openness (r_{tt} max = .62). More research is needed into the reliability of AVIs, and how (un)reliability may affect validity.

> *Research Question 1a-c*: How reliable, in terms of a) inter-algorithm, b) split-half across interview questions, and c) test-retest across occasions, are AVI ability assessments?

Convergent evidence of validity. Using empirical keying to derive predictor weights can lead to high convergence with the ground truth. As described above and in Table 2, several studies of AVIs have found strong convergence between predicted and interviewer-rated KSAOs. The present study uses this form of empirical keying to maximize the convergence between predicted and observed ability, including GMA, verbal ability, and interviewer-rated intellect. I call convergence with the ground truth measure that a model was trained to predict internal convergence because it involves the same measure used in the formula to develop the ML model.

Convergence with the ground truth measure that a model was trained to predict should be higher than the correlations with other ability measures. Further, to have practical value, such predictions should also converge with common proxies for ability, like standardized test scores and GPA—in other words, the predictions should exhibit convergence with similar measures *external* to the ML modeling process. Figure 1 summarizes the construct validation framework of the present study. Notably, although the use of ability test scores as ground truth provides a more direct linkage to the underlying construct than when self-reports or interviewer ratings are used, interviewer ratings provide a more direct linkage to interview performance than do ability test scores. I seek to examine the convergent-related validity evidence between algorithmic ability predictions, ability test scores, and proxies for ability (i.e., standardized test scores, academic performance, self-reports, and interviewer ratings).

> *Research Question 2a*: To what extent do AVI ability assessments converge with the measure of ability they were trained to predict?
>
> *Research Question 2b*: Do AVI ability assessments converge with the measure of ability they were trained to predict more strongly than they do with other measures of ability?
>
> *Research Question 2c*: To what extent do AVI ability assessments converge with commonly used proxies for ability (i.e., standardized test scores, GPA)?

Discriminant evidence of validity. Algorithmic KSAO prediction research has overwhelmingly relied on convergent evidence of validity. Yet, it is also important to demonstrate that assessments can distinguish among the various constructs they purport to assess (Campbell & Fiske, 1959). This is because empirical keying may result in poor construct discrimination (Simms, 2008). In the case of AVIs, algorithmic predictions of different constructs are generated from the same set of predictors, resulting in the same items being used to predict multiple constructs. Additionally, employment interviews tend to be contaminated with substantial method variance (Hamdani et al., 2014).

To my knowledge, only one study has provided discriminant evidence of AVI validity. Hickman, Bosch, et al. (2021) found that, compared to personality interviews (Van Iddekinge et al., 2005), AVI personality assessments exhibited lower discriminant correlations. However, even for the most accurately modeled traits, discriminant correlations sometimes exceeded convergent ones, indicating limited construct discrimination (Hickman, Bosch, et al., 2021, Appendix Tables 15 and 16). However, it is unknown whether using ability test scores as the ground truth for *ability* predictions while using interviewer-rated personality as the ground truth for *personality* predictions will help remedy this. Since the same predictors will be used to develop the algorithms, method variance may still be inflated, but since separate sources of information are used for the algorithms' ground truth in these cases, it may not. Regardless, interviewer-rated intellect predictions are likely to exhibit worse discriminant evidence than ability test score predictions. Research Question 3 regards the discriminant-related evidence of AVI ability assessments.

> *Research Question 3a*: Do AVI ability assessments exhibit expected discriminant relationships with other AVI assessments?
>
> *Research Question 3b-c*: Do AVI ability assessments exhibit expected discriminant relationships with b) self-reported and c) interviewer-rated personality traits?

Measurement bias. An additional concern for all measures, but especially for algorithmic assessments (Obermeyer et al., 2019), is bias. Bias regards systematic error in a measure that differentially affects test takers based on their group membership (SIOP, 2018), and in particular, measurement bias regards systematic error that leads members of a particular group to have inflated or deflated scores. Bias occurs because either construct irrelevant variance has contaminated a measure or because a measure is deficient, in that construct-relevant variance is not being captured. Notably, mean differences in and of themselves do not indicate bias, yet sizeable group differences should engender additional scrutiny (SIOP, 2018).

GMA tests are known to have large subgroup differences that disadvantage Black and African American test takers when compared to Whites (Cohen's d = -1), as well as moderate differences when comparing Hispanics and Whites (Cohen's d = -.5; Hough et al., 2001). Differences are also observed for verbal ability tests (Black-White d = -.6; Hispanic-White d = -.4). On the other hand, men and women tend to score approximately equal on GMA, while women score slightly higher (d = .1) on verbal ability tests.

Although mean differences in and of themselves do not indicate bias, AVIs are modeling a ground truth measure with an existing level of group differences (which may be zero, or no difference). To the extent that an ML model's predictions alter the magnitude of underlying group differences, this could be indicative of measurement bias (Tay et al., 2021). It is particularly concerning when ML models increase or *exacerbate* group differences. When this occurs, contamination or deficiency in the measure is affecting the scores and altering the relationship that group membership has to one's ability score. Therefore, I also investigate the group differences in AVI ability assessments and compare them to the observed differences in the ground truth scores.

> *Research Question 4a*: Do AVI ability assessments exacerbate group differences?

Further, since measurement bias is systematic error that disproportionately affects members of one group over another, it is also important to analyze accuracy as a function of demographics. Internal convergence, or convergent correlations between observed and predicted values, forms the basis for ML validation. Therefore, the correlations between observed and predicted values should be similar across groups (Hickman, Saef, et al., 2021)—otherwise,

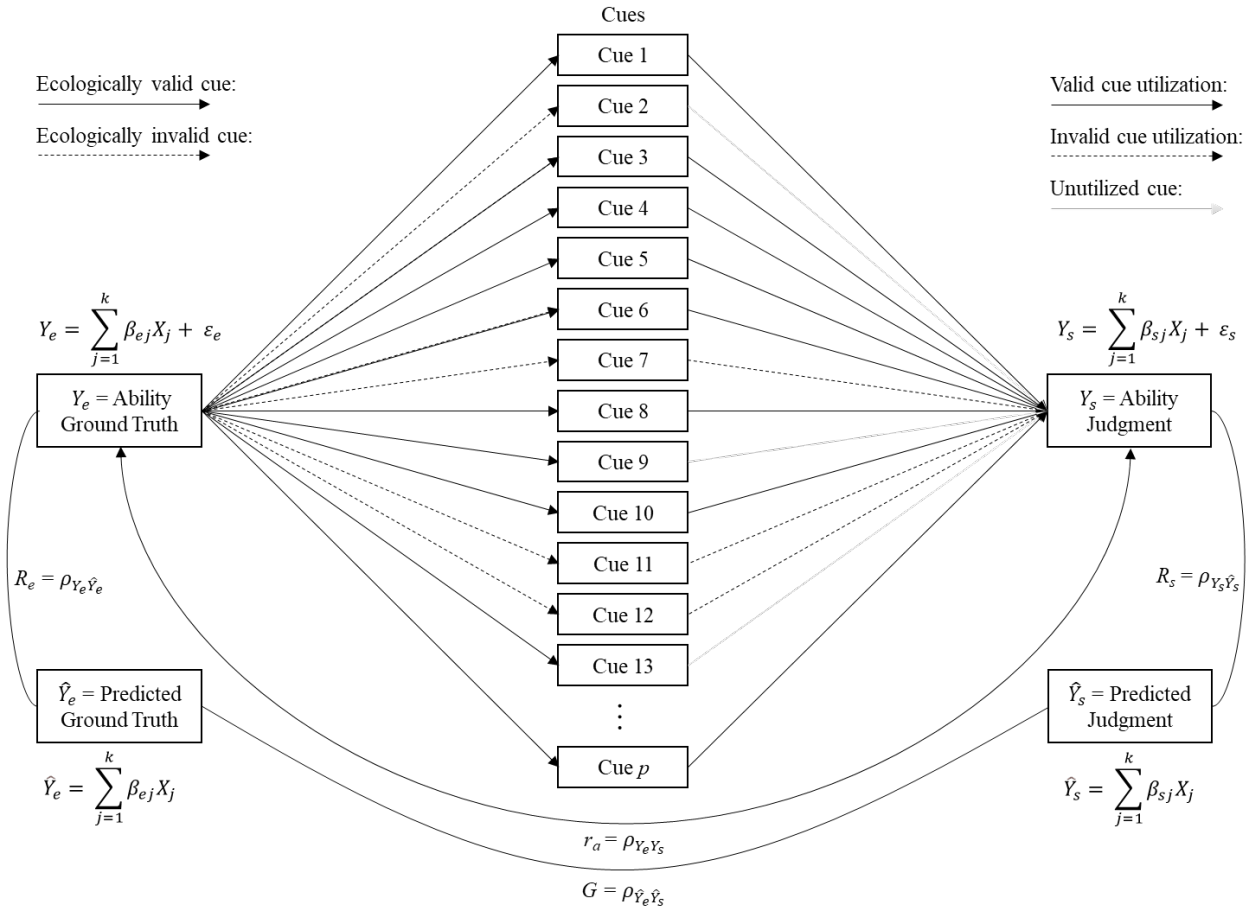
systematic error associated with group membership could inflate or deflate scores of the members of one group over another (i.e., measurement bias).

Research Question 4b: Are AVI ability assessments equally accurate across groups?

The Brunswik Lens Model

The primary focus of the present study is investigating the reliability and validity of AVI ability assessments. However, evidence of reliability and validity does not answer whether the behavioral cues used to make such assessments are conceptually relevant (i.e., content evidence of validity; SIOP, 2018). Research linking latent characteristics to specific, proximal behaviors has often adopted the Brunswik (1956) lens model. The Brunswik lens model describes how the expression of an attribute aligns with (or differs from) how others judge the attribute, as illustrated in Figure 2. The left, or environmental, side of the lens, models the relationships between a ground truth variable and behavioral cues, while the right, or judgment, side of the lens, models the relationships between behavioral cues and judgments of the ground truth variable. In the present study, the *ability ground truth* (Y_e) is either GMA or verbal ability test scores, while the *ability judgments* (Y_s) are either AVI GMA scores, AVI verbal ability scores, or interviewer ratings of intellect. Behavioral cues are the verbal, paraverbal, and nonverbal behaviors computed using text mining and computer software. In the lens model, convergence between the attribute and judgments of the attribute is termed *achievement* (r_a), and behavioral cues correlated with judgments are deemed 'valid' only if those cues also correlate with the ground truth.

The Brunswik lens is conceptually aligned with the use of supervised ML for scoring KSAOs. Brunswik lens research has long investigated the use of bootstrapping to replace human judges with a linear "model of man" (e.g., Goldberg, 1970). In such investigations, human judges are replaced with a multiple regression model that standardizes judgments by using behavioral cues to predict them (i.e., *predicted judgment* in Figure 2). In the present study, I develop multiple regression models not only for interviewer-rated intellect but also to model the AVI GMA and verbal ability predictions generated during nested *k*-fold cross-validation.



Note: Adapted from Karelaia and Hogarth (2008). Ability ground truth is either GMA or verbal ability test scores. Ability Judgment is either interviewer-rated intellect or the predictions from a GMA or verbal ability machine learning model. Cues include verbal, paraverbal, and nonverbal behaviors.

Figure 2. Brunswik lens model of ability perception.

Additionally, much ML research recognizes that many cues are interchangeable, in that using a different set of behavioral cues for prediction can lead to similar levels of model accuracy, particularly in text mining (e.g., Lee & Lee, 2006) due to high cue intercorrelations. Brunswik (1943; 1952) referred to this as vicarious mediation and vicarious functioning. Vicarious mediation occurs when cues on the environment side of the lens have high intercorrelations and are, therefore, interchangeable. Vicarious functioning regards the difficulty that this causes judges since some of the intercorrelated cues may be more valid than others.

Therefore, the Brunswik lens is an appropriate conceptual model for advancing our understanding of both human and ML ability judgments. A basic step in Brunswik lens investigations involves examining the relationships between the focal variable and the behavioral cues it causes—in the present investigation, these are the proximal behavioral manifestations of ability in the form of verbal, paraverbal, and nonverbal behavioral cues.

Proximal behavioral cues associated with ability. Although no studies have investigated the relationship between ability and proximal behaviors in employment interviews, several studies have examined the relationship between ability and proximal behaviors in other settings. Several studies have analyzed how ability affects verbal behavior. For example, Küfner et al. (2010) explored how ability related to language in a creative writing task. They found that ability related positively to ratings of writing sophistication and creativity, as well as the amount of positive emotion words used.

Pennebaker and King (1999) had participants complete a stream of consciousness writing exercise and factor analyzed LIWC scores on those texts. They found that the first factor, which they dubbed *immediacy*, was negatively related to SAT verbal scores and course exam grades. The immediacy factor included first-person singular pronouns (e.g., *I*, *me*, *my*), discrepancies (e.g., *should*, *would*), present tense verbs (a category which was removed in the most recent version of LIWC), fewer articles (e.g., *a*, *an*, and *the*), and fewer words longer than six letters.

In a similar vein, Pennebaker et al. (2014) analyzed how word usage in college admissions essays related to standardized test scores (i.e., SAT or ACT scores) and college GPA. They argued that more categorical language consists of more abstract thinking, as reflected in the use of more articles, and greater cognitive complexity, as reflected in the use of more prepositions (e.g., *to*, *with*, *above*), while dynamic language consists of a more narrative language style, as reflected in

the use of more adverbs (e.g., *very, really*), auxiliary verbs (e.g., *am, will, have*), impersonal pronouns (e.g., *it, those*), personal pronouns (e.g., *I, them, her*), conjunctions (e.g., *and, but*), and negations (e.g., *no, never*). The combination of these eight categories forms the analytical thinking LIWC category, with higher scores indicating more categorical language. Categorical language was positively correlated with both standardized test scores and college GPA. Pennebaker et al. (2014) also examined which specific elements of analytical thinking were related to college GPA, finding that each element was related in the expected direction, with the weakest relationships observed for prepositions and negations.

Robinson et al. (2013) examined how written self-introductions related to final college course performance. Numerous LIWC and non-LIWC variables were related to final grades, including positive relationships for readability, word count, first person plural pronouns (e.g., *we, us, our*), and certainty (e.g., *always, never*); and negative relationships for personal pronouns, first person singular pronouns (e.g., *I, me, mine*), common verbs (e.g., *eat, come, carry*), auxiliary verbs, friend (e.g., *buddy, neighbor*), and home (e.g., *kitchen, landlord*). Overall, ability appears to have some consistent relationships with word usage, including negative relationships with some types of pronouns and auxiliary verbs, and positive relationships with articles and words longer than six letters.

Little theory has connected nonverbal and paraverbal behaviors directly to ability, yet some related works can be insightful. Regarding paraverbal behavior, ability affords several benefits. Individuals with higher ability tend to speak for a longer duration (Murphy, 2007), with a higher speech rate, and are easier to understand (Borkenau & Liebler, 1995; Reynolds & Gifford, 2001). Each of these paraverbal behaviors is positively related to interviewer ratings (DeGroot & Motowidlo, 1999; Feiler & Powell, 2016).

Regarding nonverbal behavior, Darwin (1872) asserted that extended periods of concentration are often accompanied by frowns. So, people with lower ability may be more likely to frown during extended periods of concentration and self-regulation, such as occur during employment interviews. A study of computer science students found that automatic measurements of action unit 17, the *mentalis* or chin raiser, which is visually similar to a frown, is common during coding sessions (Tiam-Lee & Sumi, 2017). One possibility is that interviewees with lower levels of ability were involuntarily activating their mentalis (chin raiser) muscle. Combinations of the chin raiser with action unit 15, the lip corner depressor, are present during expressions of doubt

and uncertainty (Bitti et al., 2014), feelings which may be persistent for interviewees with less ability to maintain concentration and self-control during the interview. Action unit 14, *buccinator*, or the dimpler, may also be relevant as it is associated with feelings of anxiety and discomfort (Ozel, n.d.).

Although several works have examined how ability and proxies for it relate to proximal verbal, paraverbal, and nonverbal behaviors, it remains to be seen whether these relationships are also evident in employment interviews.

*Research Question 5*a: How does ability relate to interviewee behavior?

Behavioral cues used to infer ability. Like the lack of knowledge regarding the environmental side of the lens for interviewee ability, there is a corresponding lack of knowledge regarding the judgmental side of the lens. However, understanding the judgmental side of the lens is important to understand whether interviewers rely on valid cues, or if irrelevant cues, like attractiveness, are used to judge ability. The validity of cues that interviewers use to judge ability has not been examined, yet some research has investigated the validity of such cues in other situations.

On average, observers judge ability about as accurately as people self-report it—the meta-analytic accuracy of observer-rated ability is $r = .30$ (Zebrowitz et al., 2002). However, observers often use several invalid cues to judge ability. For instance, ability ratings are strongly related to attractiveness, especially for female targets (Borkenau & Liebler, 1995; Kleisner et al., 2014). However, among adults, ability is unrelated to attractiveness (Kleisner et al., 2014; Lee et al., 2017; Zebrowitz et al., 2002). Similarly, attractiveness is related to interview ratings but not job performance (Barrick, Shaffer, & DeGrassi, 2009). Other superficial characteristics, like stuttering and shyness, are also related to ability judgments but are not valid cues (Paulhus & Morgan, 1997; Zeigler-Hill et al., 2019).

Observers do tend to use some valid cues as well. For instance, people are judged to have higher ability when they say more words, speak at a faster rate, and have pleasant sounding voices that are clear and easy to understand (Borkenau & Liebler, 1995; Murphy, 2007; Murphy et al., 2003; Reynolds & Gifford, 2001). When judging intelligence from writing, the sophistication,

creativity, and positivity of the writing are all positively related to ability judgments (Küfner et al., 2010).

Another element that may contribute to inaccuracy in ability judgments is inconsistency in how judgments are made. For example, sociocognitive, racial, and ethnic biases may contaminate judgments for some targets, such as occurs when attractiveness is disproportionately used to judge female's ability (e.g., Borkenau & Liebler, 1995). Fatigue can also reduce one's ability to notice behavioral cues. The Brunswik lens model equations (Hursch et al., 1964; Tucker, 1964) help to capture some of these effects on judgment accuracy, are presented in Figure 2, and are described below[1].

First, the Brunswik lens analysis involves creating a linear model of each side of the lens using ordinary least squares regression. The *ability ground truth* scores (in this case, either GMA or verbal ability test scores), Y_e, are regressed onto a portion of the verbal, paraverbal, and nonverbal behavioral cues (see: Table 1), X_j, where j = 1, …, k (and k = the number of behavioral cues). This represents the environmental side of the lens. Specifically:

$$Y_e = \sum_{j=1}^{k} \beta_{ej} X_j + \varepsilon_e \tag{1}$$

Formula 1 can be used to generate linearly *predicted ground truth*:

$$\hat{Y}_e = \sum_{j=1}^{k} \beta_{ej} X_j \tag{2}$$

Separately, the *ability judgments* (in this case, AVI GMA scores, AVI verbal ability scores, or interviewer-rated intellect), Y_s, are regressed onto the behavioral cues, representing the judgment side of the lens:

[1] The present study is unique in that over 5,000 behavioral cues were available for analysis across all modalities. All behavioral cues were considered when exploring cue ecology and utilization, but a subset of approximately 250 behavioral cues that were most strongly correlated with GMA and verbal ability, respectively, were used when calculating Brunswik lens model equations.

$$Y_s = \sum_{j=1}^{k} \beta_{sj} X_j + \varepsilon_s \quad (3)$$

And formula 3 can then be used to generate linearly *predicted judgments*:

$$\hat{Y}_s = \sum_{j=1}^{k} \beta_{sj} X_j \quad (4)$$

$\hat{Y}_e$, $\hat{Y}_s$, ε_e, and ε_s are then used to decompose the achievement index, r_a, which is the correlation between the ground truth, Y_e, and judgments, Y_s. Specifically:

$$r_a = \rho_{Y_e Y_s} = G R_e R_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)} \quad (5)$$

G is the matching index, or knowledge, which is calculated by correlating the predictions from formulae 2 and 4 above, or $\rho_{\hat{Y}_e \hat{Y}_s}$. G expresses the extent to which the behaviors used to judge the construct correspond to the behaviors expressed by the latent construct—the extent to which each cue is used relative to its validity. Higher values indicate that the judge used cues more validly.

R_e is the multiple correlation between the ability ground truth, Y_e, and linearly predicted ground truth, $\hat{Y}_e$, or $\rho_{Y_e \hat{Y}_e}$. It represents the upper limit of environmental predictability. This is relevant to the ML context in that R_e provides an upper limit of the predictability of the ground truth *without* cross-validation, and therefore, cross-validated predictions from AVIs are unlikely to ever exceed this value.

R_s is the multiple correlation between the ability judgments, Y_s, and linearly predicted judgments, $\hat{Y}_s$, or $\rho_{Y_s \hat{Y}_s}$. It represents the consistency with which judgments are made across targets, known as judgmental consistency. When considering interviewer-rated intellect, R_s relates to ML similarly to R_e, in that it represents an upper limit on the predictability of ratings without cross-validation. When the judge is a series of AVI GMA or verbal ability models, such as trained and tested during nested *k*-fold cross-validation, R_s represents the consistency of weights assigned to cues across the *k* models.

C is the correlation between the two regression models' residuals, ε_e and ε_s. When C is non-zero, it suggests that relevant cues may have been omitted from one or both of the models, nonlinear relationships between behavioral cues and Y_e or Y_s, or a combination of the two.

Additionally, two composite indices are formed by taking the product of G and judgmental consistency, R_s, and environmental predictability, R_e. GR_s estimates the judge's contribution to achievement and is known as performance. It indicates both how well judges matched task requirements and how consistent they were in making judgments. GR_e estimates the validity of a linear model that is created by replacing the judge with their strategy, which estimates the achievement that would occur if the judgments were made in a completely consistent manner across targets (i.e., when $R_s = 1$). For interviewer-rated intellect, behavioral cue-ability judgments may be inconsistent due to sociocognitive biases and fatigue. For ML models, behavior-ability judgments are only inconsistent when resampling methods that involve multiple train/test splits, such as *k*-fold cross-validation, are used. This is true even when ML models are trained on biased human ratings because ML modeling involves creating a single model that is applied to make all subsequent judgments. The extent of consistency across the *k* models speaks to the robustness of behavior-ability relationships across resampling iterations.

Each of these Brunswik lens variables speak to the relative advantages of one judgment method (i.e., AVI or interviewer-rating) over another. Additionally, the visual component of the Brunswik lens model involves illustrating which behavioral cues are (in)valid, and which (in)valid cues are related to ability judgments. As mentioned, interviewer-rated ability is a worse predictor of job performance than ability test scores (Huffcutt et al., 2001). Using the Brunswik lens model to compare how ability is expressed behaviorally (i.e., the valid cues) versus the cues interviewers use to judge ability may help elucidate why. For example, using longer words may be a valid cue but may not be utilized by interviewers, or invalid cues like voice pitch may be related to ability ratings. In the parlance of the lens model, algorithms may use more valid cues more consistently than interviewers to judge ability because the algorithms are trained on ability test scores.

> *Research Question 5b-e*: What cues do b) interviewers and c) algorithms use to judge interviewee ability, d) who uses cues more validly, and e) who uses them more consistently?

METHODS

Participants and Procedure

Sample 1

I recruited 774 non-freshman undergraduate students to participate in the study in exchange for a $10 Amazon gift card. These students were recruited from a variety of sources and universities, including via direct email, posting to university study lists, and the online panel service Prolific (where participants were instead compensated with a direct payment of $7.20). The study was administered online and consisted of a series of common selection tests, including a self-reported personality test, two ability tests (GMA and verbal ability), and a mock asynchronous video interview. Participants were encouraged to use the study to gain interview experience and practice their skills, as several prior studies have done (e.g., Van Iddekinge et al., 2005). After extracting verbal, paraverbal, and nonverbal behaviors from the videos, 47 participants were removed due to missing features (due to issues with their video submissions), leaving a final sample $N = 733$.

Sample 2

I recruited 226 psychology subject pool participants to complete the study in exchange for course credit. This sample participated twice in the mock video interview, with the second interview occurring 4 to 47 days after the first (median 8 days, mean 9.2 days). At Time 1, the study was identical to the study completed by Sample 1. At Time 2, participants completed the GMA test, verbal ability test, and the mock video interview again. Of the 226 participants, 25 had missing features for at least one time point, leaving a final sample $N = 201$.

Measures

General mental ability. General mental ability was assessed using a 16-item test from The International Cognitive Ability Resource (ICAR, 2014). The test consisted of four three-dimensional rotation questions, four letter series questions, four matrix completion questions, and four verbal reasoning questions. The items were presented in random order. The items do not require specific domain knowledge (beyond basic algebra and English skills) so they represent an appropriate test of GMA. Participants were given 12 minutes to answer as many questions as possible. Reliability for this test and all other measures is provided in the diagonal of Tables 3 and 4. Cronbach's alpha for this test was .70 in Sample 1, and .60 and .76 in Sample 2 at Times 1 and 2, respectively. Additionally, the test-retest reliability of this test in Sample 2 was .66. The scores were approximately normally distributed in Sample 1 (Appendix Figure 9). The internal consistency of the test may have been attenuated due to unevenly distributed missingness. In Sample 1, on average across the four three-dimensional rotation questions, 13.5% of participants did not select an answer, and on average across the 12 remaining questions, 6.3% of participants did not select an answer. Validity evidence for the GMA and verbal ability tests are provided in the first Results section.

Verbal ability. Verbal ability was assessed using a custom-developed 19-question multiple-choice test designed to be like questions in the verbal portion of the GRE. An initial item bank was developed consisting of 22 items that were reviewed and pilot tested by several undergraduate researchers and industrial-organizational psychology faculty. Three items were eliminated upon this review, leaving the 19 items used in the study. Participants were given 12 minutes to answer as many questions as possible. In the first type of items, two answers are selected to complete a blank in the sentence. A sample item is, "Some social commentators have labeled bankers as _____ for their role in causing the 2008 financial crisis." Answer options included: avaricious; rapacious; treacherous; impenitent; insensate; solicitous. In the second item type, one answer is selected to complete each of one to three blanks in a sentence. A sample item is, "The immigrant's poor English skills are hardly an ____ problem; he can attend classes and improve within a few months." Answer options included: insuperable; implausible; inconsequential; evocative; injudicious. The test consisted of nine of the first type of question and 10 of the second

type of question. Within each type of question, the questions were presented in random order. Participants were rewarded one point for each correct response in each question, with a maximum possible score of 37. Cronbach's alpha for this test was .84 in Sample 1, and .79 and .81 at times 1 and 2 in Sample 2. Additionally, the test-retest reliability of this test in Sample 2 was .80. For comparison, Raven's standard progressive matrices have test-retest reliability of .78 (Kosinski et al., 2013). The scores were approximately normally distributed in Sample 1 (Appendix Figure 10). In Sample 1, on average across the first set of questions, 0.7% of participants did not select an answer, and on average across the second set of questions (which were presented second), 5.3% of participants did not select an answer.

Self-reported GMA. Self-reported GMA was measured using two items, both of which were relative scales and one of which provided a reference group, per recommendations by Freund and Kasten (2012). Specifically, the first item provided an illustration of a bell curve and explained the IQ distribution (e.g., IQ of 100 is average intelligence) in one standard deviation intervals, adapted from Chamorro-Premuzic and Furnham (2006), and asks respondents to rate their own IQ. The second item asked participants to, "On the slider below, please rate your intelligence percentile compared to the average student at Purdue University (e.g., 5^{th} percentile indicates you are more intelligent than 5% of Purdue students; 95^{th} percentile indicates you are more intelligent than 95% of Purdue students.)" Once recruitment expanded to other universities, the second item was amended to read "your university." However, upon visual inspection of the histogram, responses to the second item were highly skewed left, and the second item correlated minimally with ability test scores, so it was dropped from further analysis.

Self-reported personality. Five factor model (FFM) traits were assessed using the 44-item Big Five Inventory (BFI; John & Srivastava, 1999). The scale was adapted by asking participants to respond how they typically act at work. Participants indicated the extent to which each statement describes them via a five-point Likert scale (ranging from Disagree strongly to Agree strongly). An example item for Extraversion is, "Has an assertive personality." An example item for Agreeableness is, "Is sometimes rude to others." An example item for Conscientiousness is, "Perseveres until the task is finished." An example item from Emotional Stability is, "Is related,

handles stress well." An example item for Openness is, "Is inventive." Cronbach's alpha ranged from .77 (agreeableness and openness) to .88 (extraversion).

Proxies for ability. Participants in both samples self-reported their college GPA. Additionally, participants were asked whether they took the SAT, ACT, or both, and were asked to report their SAT verbal, SAT math, and ACT scores if they took the respective tests. Such self-reports converge highly ($r > .8$) with the actual scores (Kuncel et al., 2005). Students who were in their first semester of college had their college GPA disregarded, and college GPA was disregarded in Sample 2 because most of these participants were in their first semester.

Attention check. An attention check item was included in the personality inventory. Participants who failed to answer it correctly had their self-reports dropped from further analysis (Meade & Craig, 2012). In Sample 1, 10 participants self-reports were dropped (leaving $N = 723$), and in Sample 2, all 201 of the retained participants passed the attention check.

Mock video interview. The mock video interview consisted of five past behavior and one situational interview question. The six questions were developed such that one question taps each of the FFM traits and ability. Participants were instructed to answer each question for 1-3 minutes. The interview is embedded into the online survey, and participants were first provided with a chance to familiarize themselves with the web-based recorder by recording their answer to the prompt, "Tell us about your dream job." Participants were encouraged to take time to prepare their responses by reflecting on their past work experiences and accomplishments, and if they did not have relevant work experiences, to think of experiences from school, volunteering, or other organized activities. The six interview questions were presented in random order. The question corresponding to Extraversion is, "Do you prefer to work alone or in a team? Tell us about a time you had to work against your preference (e.g., had to work in a team when you prefer to work alone). What were the challenges you faced and were you able to overcome them?" The question corresponding to Agreeableness is, "Think of a time a coworker asked you to set aside your own work to help him or her with a project that was very important to them. What did you do? Why did you do that?" The question corresponding to Conscientiousness is, "Describe a long-term project that you managed. What did you do to keep everything moving along in a timely manner?" The question corresponding to Emotional Stability is, "Tell me about a recent uncomfortable or

difficult work situation. How did you approach this situation? What happened?" The question corresponding to Openness is, "Think of a time you had a need to learn about something that was new to you? Why did you pursue it? What kept you persistent?" The question corresponding to ability is, "You have been tasked on your job to make product purchases for the company. Explain step-by-step how you would choose between two or three different products." Interviews were retained if at least four videos were usable for analysis. In Sample 1, the responses averaged 1308.6 words across the entire interview. In Sample 2, the responses averaged 1279.2 words across the entire interview at Time 1 and 1163.2 words at Time 2.

Interviewer-rated personality. From a pool of 12 undergraduate research assistants who underwent one to two hours of frame-of-reference training, four watched each interview in Sample 1 and rated each participant's personality. The 'interviewers' were paid $8 per hour. Interviewers rated interviewee personality on a seven-point Likert scale using an observer version of the Ten Item Personality Inventory (Gosling et al., 2003). An example item for Extraversion is, "Extraverted, enthusiastic." The average of the four trait estimates was used as the final interviewer-reported traits. A small proportion of participants (15%) received only three ratings. Average one-way random effects intraclass correlations ICC(1, k) ranged from .57 (openness) to .75 (extraversion).

Interviewer-rated intellect. The same undergraduate research assistants also rated each Sample 1 participant's intellect (sometimes considered a facet of openness) on a seven-point Likert scale using two items, "intelligent, bright" and "has a good vocabulary," adapted from Kluemper et al. (2015). The ICC(1, k) for these ratings was .61.

Interviewer-rated hireability. The same undergraduate research assistants also rated each Sample 1 participant's hireability on a five-point Likert scale. The research assistants were instructed to consider the interviewee's suitability for a managerial or team lead role. They responded to two items, "I would recommend that this person be hired" and "If hired, I believe this person would perform well on the job," adapted from Dunn et al. (1995). The ICC(1, k) for these ratings was .75.

Verbal behavior. To analyze interviewee verbal behavior, the interviews were first transcribed using IBM Watson Speech-to-Text. Verbal behavior was then operationalized in multiple ways. First, I scored each question's transcript using all Linguistic Inquiry and Word Count (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015) variables, then calculated both the mean across the questions and the standard deviation of these scores. LIWC scores a series of semantic and psychological dictionaries based on the proportion of text represented by those dictionaries.

Second, using the interview-level transcripts, I used the quanteda R package (Benoit et al., 2018) to calculate the lexical diversity and readability of the speech, as well as the average number of syllables in each word. To calculate lexical diversity, I used Guiraud's *Root TTR* (1954), which is calculated by dividing the unique tokens by the square root of the total number of words. To calculate readability, I needed an index that did not rely on sentence or paragraph length, since there is no punctuation in the transcripts. Coleman's (1971) Readability Formula 1 and FORCAST (Caylor & Sticht, 1973) are mathematically equivalent and in prior work correlate r = .90 with Flesch's Reading Ease Score. The FORCAST formula is simpler than Coleman's (1971) formula:

$$\text{FORCAST} = 20 - \frac{\left(N_{wsy1} * 150\right)}{\left(N_w * 10\right)} \tag{6}$$

N_{wsy1} is the number of one-syllable words, and N_w is the total number of words.

Third, using the interview-level transcripts, I used the tm R package (Feinerer & Hornik, 2019) to extract *n*-grams where n = 1, 2, and 3 (i.e., unigrams, bigrams, and trigrams, or one-, two-, and three-word phrases, respectively). Following recent recommendations (Hickman et al., 2020), the text was preprocessed by first appending a series of negation terms (i.e., not, n't, cannot, never, no) to the subsequent words. Then, all numbers were removed, all text was converted to lowercase, all punctuation was removed, and all words were stemmed. Before extracting unigrams, stop words were removed, but stop words were not removed before counting bigrams and trigrams to avoid the creation of nonsensical phrases. If an *n*-gram did not occur in at least 5% of the documents in Sample 1, it was removed. The raw count of each *n*-gram was used for subsequent analysis.

Fourth, using the topicmodels R package (Hornik & Grün, 2011), I generated Latent Dirichlet Allocation (LDA) topic models that assign both words and documents (i.e., interview

transcripts) scores that indicate the extent to which a given topic is represented by that word or is present in each document (Blei et al., 2003). I used the ldatuning R package (Murzintcev & Chaney, 2015) to explore the appropriate number of topics for Sample 1's interview transcripts. The ldatuning package provides four statistics that suggest how many topics are appropriate in a dataset. The statistic described by Deveaud et al. (2014) was the only one with a clear curve—it reached its highest values at 30, 40, and 50 topics, then began decreasing with larger numbers of topics. This, coupled with a visible elbow in the Cao et al. (2009) statistic at 50 topics, suggested that 50 topics was optimal. These 50 topics were then extracted, and each document was described by the extent to which each topic was present in the text (with the sum of the 50 topics equal to one for each document). The top 10 *n*-grams in each topic are provided in Figure 3.

Fifth, I applied DistilBERT to the question-level transcripts using the transformers Python package (Wolf et al., 2020). DistilBERT distills BERT, the transformer-based transfer learning language model, into a set of 768 parameters that represent the semantics of text (Sanh et al., 2019). BERT relies on deep learning both for language representation and downstream tasks (e.g., supervised ML), making it computationally intensive and cumbersome to use, particularly when applied to nested cross-validation (where the deep learning tuning and training would need to occur separately within each training fold). DistilBERT requires far fewer computational resources, outputs parameters that can be used in traditional (i.e., non-deep learning) ML pipelines, and retains 97% of the performance of BERT on the General Language Understanding Evaluation task (Sanh et al., 2019). DistilBERT had to be applied at the question-level because it has a token limit of 512 (DistilGPT-2 has a token limit of 1024, but this would still have been inadequate because the average interview-level word count in Sample 1 was 1,308.6). After generating DistilBERT parameters for each question, I then calculated the means and standard deviations of each parameter and used those for modeling.

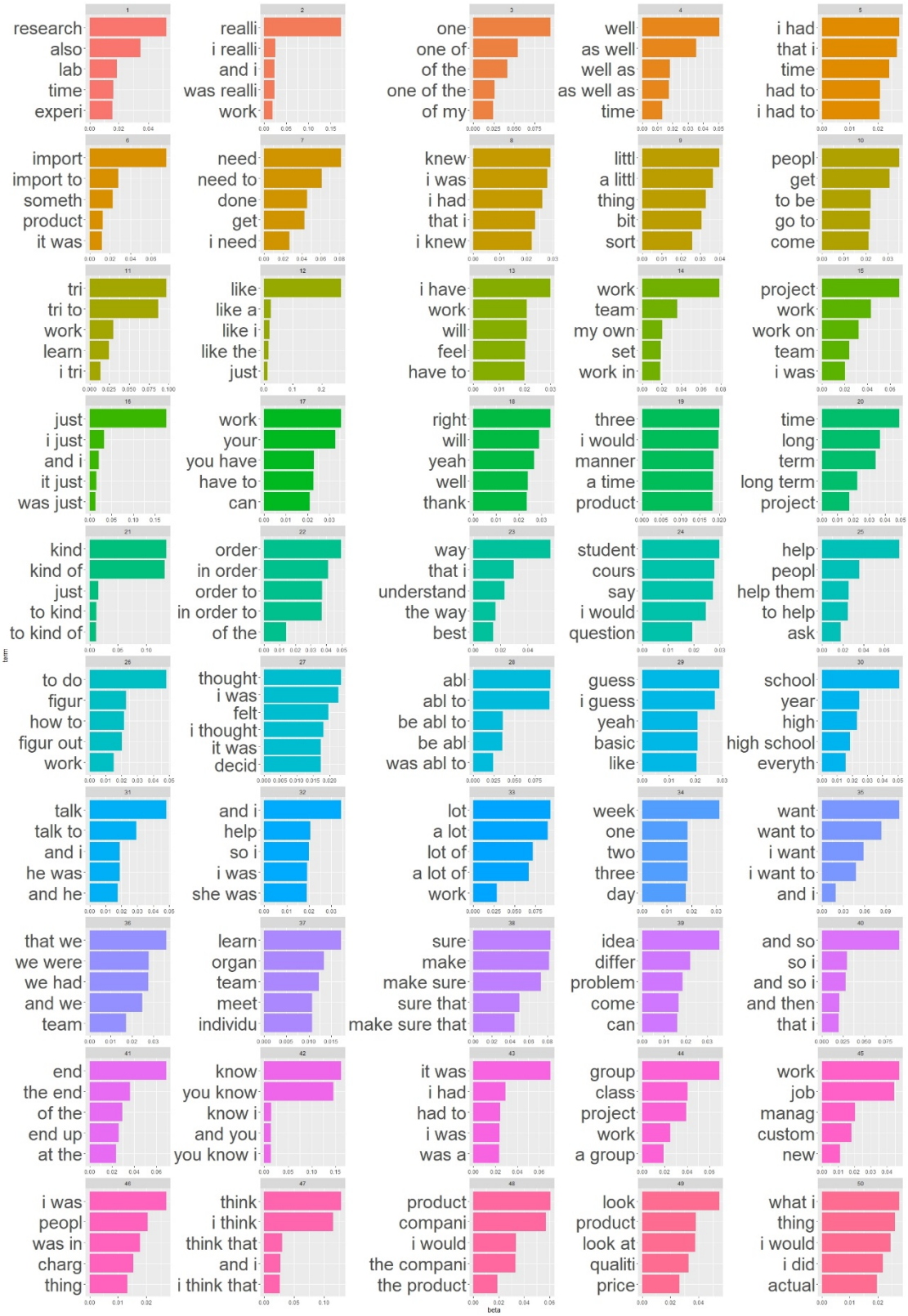


Figure 3. Top 5 terms in Latent Dirichlet Allocation topics.

Paraverbal behavior. Paraverbal behaviors were quantified using openSMILE to extract the Geneva Minimalistic Acoustic Parameter Set (Eyben, 2014; Eyben et al., 2016). These audio features were extracted from 30-second sliding windows of voice data within each interview question. The features were then summarized across the interview by calculating their means and standard deviations. Additionally, paraverbal behavior was described by calculating the number of utterances, the average duration of pauses (by diving the sum of time between utterances by the number of utterances), the number of stop words per second, and the number of filler words per second and again calculating their mean and standard deviation across videos.

Nonverbal behavior. Nonverbal behaviors were quantified using Emotient (2015), a software library for facial expression recognition. Emotient extracts information about 20 facial action units (Ekman & Friesen, 1978), seven basic emotions (i.e., anger, contempt, disgust, joy, fear, sadness, and surprise), and information about head pose along horizontal (pitch), vertical (yaw), and depth (roll) dimensions. The video time series data was summarized by calculating the mean and standard deviation of each feature.

Algorithmic Assessments

Ground truth. The focal predictive algorithms were trained to model several ability assessments, including GMA test scores, verbal ability test scores, self-reported intelligence, and interviewer-rated intellect. Additionally, to investigate discriminant evidence of validity, six additional algorithms were trained to model interviewer-rated Big Five traits and hireability. These 10 variables, then, formed the ground truth for the ML models.

Cross-validation strategy. In Sample 1, all models were developed and tested using 10-fold nested cross-validation with $k = 5$ inner folds. In this procedure, the data is first split into 10 equally sized parts, and nine of the parts (the outer training folds) are used initially to conduct an inner 5-fold cross-validation to identify the optimal model hyperparameters. Then, a model is trained on those nine parts with the optimal hyperparameters. In each set of training data, I discarded all predictors that correlated $|r| < .03$ with the focal outcome to reduce the n:p ratio and the odds of overfitting. The model trained on those nine parts of the data was then used to predict the focal outcome in the tenth fold, and that process was repeated 10 times, using each outer fold only once for testing.

I conducted this process separately for each type of predictor extracted. Specifically, I created models using: 1) LIWC, lexical diversity, and readability measures (hereafter, *LIWC Plus*); 2) *n*-grams; 3) topic models; 4) DistilBERT; 5) nonverbal behaviors; and 6) paraverbal behaviors. Additionally, I created a model using 7) LIWC Plus and *n*-grams together as predictors (hereafter, the *combination model*), since similar models have performed well in prior research (Hickman, Bosch, et al., 2021).

To estimate the accuracy of the ML models, the cross-validated predictions were correlated with the outcome the model was trained to predict (RQ 2a), the other ability measures (RQ 2b), self-reported GPA and standardized test scores (RQ 2c), algorithmically assessed personality (RQ 3a), and self-reported and interviewer-rated personality (RQ 3b-c). The cross-validation process is illustrated in Figure 4 and the overall validation process in Figure 1.

To investigate RQ 4a, I calculated the means and standard deviations by ethnicity and gender for both observed ability scores and the combination models' predicted values. For RQ 4b, I calculated convergence with the ground truth measure by ethnicity and gender.

To generate models for assessing Sample 2 interviewees, a thrice-repeated 10-fold cross-validation was conducted on Sample 1 to identify optimal hyperparameters. Then, a model was trained on the full set of Sample 1 participants using those optimal hyperparameters. That model was then applied to Sample 2 to assess the focal outcome. Research questions 2a-c were also examined in Sample 2 using the average AVI scores from Times 1 and 2.

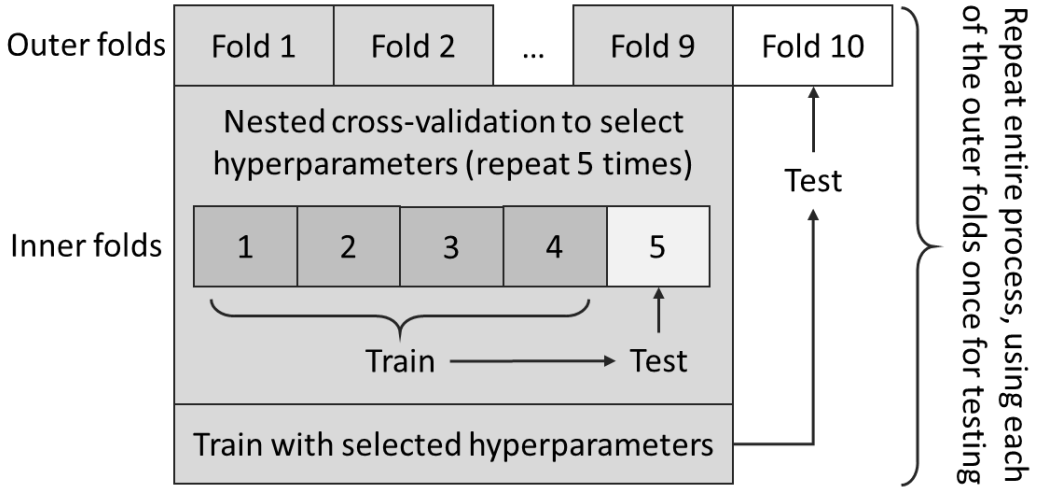


Figure 4. Nested cross-validation procedure in present study.

Predictive algorithms and inter-algorithm reliability. Multiple algorithms were trained and estimated for two purposes: 1) to identify which algorithm provides the most accurate estimates of ability, and 2) to calculate inter-algorithm reliability (r_{aa}; RQ 1a). Specifically, elastic net regression and random forest were used. Elastic net regression is a hybrid of ridge regression and least absolute squares shrinkage (LASSO) regression (Zou & Hastie, 2005). It combines the benefits of ridge and LASSO, penalizing beta weights for multicollinearity as in ridge regression, and removing some predictors from the model due to model complexity as in LASSO. Hyperparameter tuning provides the optimal amount of a) penalization and b) balance between ridge and LASSO. In each instance of inner cross-validation, I tried 10 values of lambda, which determines how severely regression weights are penalized, and these values were generated by caret then held constant across all instances. I tried 11 values of alpha, which determines whether elastic net acts more like ridge of LASSO, ranging from 0 (ridge) to 1 (LASSO) stepping by .1.

Random forest is an ensemble learning model that combines the predictions of multiple decision trees (Breiman, 2001). In each tree, only a subset of predictors and participants is used to make predictions. I tuned hyperparameters that determine the number of trees in the forest and the number of predictors given to each tree. For the number of trees, I tried 50 to 350, in steps of 50. For the number of predictors, I tried log(p), $\sqrt{p}$, p/2, and p (cf. Gladstone et al., 2019), where p = the number of predictors retained after eliminating predictors correlated $|r| < .03$ with the focal outcome. By using different predictors and observations in each tree, the model tends to be robust to overfitting and often has high accuracy because each tree uses different information. Inter-algorithm reliability was calculated in Sample 1 by correlating the cross-validated predictions from elastic net and random forest.

Split-half reliability. To examine split-half reliability (RQ 1b), the interviewee response in Sample 2 at Time 1 was split in half using an odd-even question split. Each feature was then calculated separately on these halves of the interviewee response, and algorithmic predictions of ability were generated for each half and correlated. The odd numbered question responses averaged 654/95 words, and the even numbered questions responses averaged 624.03 words.

Test-retest reliability. To examine test-retest reliability (RQ 1c), assessments were made on Sample 2 participants at Times 1 and 2. Those assessments were then correlated. The average of these two assessments was also used to examine RQ 2.

Construct convergence and discrimination. To quantify the discriminant evidence of validity and method variance (RQ 3), MTMM matrices were used to calculate Woehr et al.'s (2012) convergence, discrimination, and method variance indices. Estimates calculated directly from MTMM matrices converge with the results from more complex analyses, including confirmatory factor analytic techniques.

Brunswik Lens Model

To examine the differences in how ability is expressed versus how it is judged, the Brunswik lens model (1956) provides a theoretical and empirical underpinning (RQ 5a-e). In the Brunswik lens model, the ability ground truth (in this case, GMA and verbal ability test scores) is related to behavioral cues (i.e., verbal, paraverbal, and nonverbal cues), with significant correlations indicating the 'valid' cues for that variable. Then, a parallel procedure is conducted to examine the cues used by observers to judge ability. If ability judgments are correlated with behavioral cues that are not correlated with the ability ground truth, it indicates that judges (i.e., either AVI GMA models, AVI verbal ability models, or interviewers) may be using irrelevant information to make judgments. In the Brunswik lens model, the correlation between ability ground truth and ability judgments is dubbed the "achievement index." For analyzing cue validity and utilization, I considered all available cues across modalities.

Additionally, the Brunswik lens model equations described above express the consistency between the behavioral expression of a characteristic and the judgment of that characteristic. In the present study, although ML models were made with specific subsets of behavioral cues, the lens model equations were analyzed using all types of behavioral cues (i.e., verbal, paraverbal, and nonverbal). To select a subset of cues for analysis, I limited the calculation of lens model equations to approximately 250 predictors that were most strongly correlated with GMA and verbal ability, respectively. Including all available cues was impractical because R^2 equals 1 when too many predictors are included in the linear prediction models, and including too few predictors was untenable because the judgments were relevant to many cues. Each element of the lens model

equations speaks to how behavioral cues are related to ability, how behavioral cues relate to judgments, or how judgments relate to ability.

RESULTS

The research questions, how they are tested, and key results are provided in Table 5.

Descriptive Statistics and Ability Test Validity

Table 3 presents the descriptive statistics and intercorrelations of the variables for Sample 1, and Table 4 presents that information for Sample 2. In Sample 1, contrary to prior research (e.g., Hickman, Bosch, et al., 2021), the personality self-reports and interviewer ratings were relatively independent. The convergent correlations for extraversion, conscientiousness, and emotional stability were statistically significant (*p*s = .04, .004, & .005, respectively) but small in magnitude (*r*s = .08, .11, & .10, respectively).

Meta-analyses estimate that self-reports converge *r* = .33 with tested ability in low stakes settings (Freund & Kasten, 2012), and observer judgments converge *r* = .30 (Zebrowitz et al., 2002). In Sample 1 (Table 3), self-reported IQ was moderately and significantly positively correlated with all proxies for and direct measures of ability, with *r*s ranging from .30 (self-reported SAT verbal scores and GMA test scores) to .40 (self-reported ACT scores). Additionally, self-reported IQ was positively correlated with college GPA (*r* = .18). Self-reported openness was also positively correlated with verbal ability test scores (*r* = .08).

Compared to self-reported IQ, interviewer rated intellect tended to be slightly less strongly correlated with the proxies for and direct measures of ability, with *r*s ranging from .16 (self-reported SAT verbal scores) to .33 (Verbal ability test scores). Additionally, interviewer rated intellect was positively correlated with college GPA (*r* = .13). Self-reported IQ and interviewer-rated intellect were also positively correlated (*r* = .18). Compared to intellect ratings, interviewer ratings of conscientiousness had larger, but not significantly different, correlations with self-reported SAT verbal scores and college GPA (*r*s were .01 & .03 larger, respectively). Otherwise, interviewer-rated intellect had stronger correlations with these direct measures of and proxies for ability than did any other interviewer rated construct.

Table 3. Correlation Matrix of Observed Variables in Sample 1

	Mean	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Gender	.64	.48																					
2. Age	21.73	3.88	-.01																				
3. Self-reported IQ	113.24	10.03	-.12	-.05																			
4. SAT verbal	654.76	96.86	.04	-.08	.30																		
5. SAT math	651.66	116.40	-.15	-.13	.34	.56																	
6. ACT	29.03	4.49	-.10	-.25	.40	.57	.63																
7. College GPA	3.40	.72	.07	.02	.18	.15	.17	.21															
8. GMA	8.19	3.23	-.03	-.08	.30	.31	.36	.42	.13	(.70)													
9. Verbal ability	19.10	7.59	.03	-.04	.31	.47	.31	.52	.16	.43	(.84)												
Self-Reports																							
10. Extraversion	3.29	.86	.00	-.01	-.05	-.11	-.05	-.01	-.01	.02	-.08	(.88)											
11. Agreeableness	3.93	.60	.03	-.01	.00	.01	.02	.03	-.01	.01	-.04	.15	(.77)										
12. Conscientiousness	3.84	.67	.04	-.01	.03	-.01	.05	.07	.04	.07	.07	.12	.38	(.83)									
13. Emotional Stability	3.17	.80	-.03	.01	.01	-.04	-.02	.00	.01	.09	.00	.32	.30	.40	(.84)								
14. Openness	3.71	.59	-.01	-.03	.01	.09	.04	.02	-.04	.06	.08	.27	.18	.03	.02	(.77)							
Interviewer Ratings																							
15. Intellect	5.47	.67	.01	-.04	.18	.16	.21	.32	.13	.24	.33	.05	.02	.11	.10	-.01	(.61)						
16. Extraversion	4.45	1.14	.07	-.11	.13	.06	.04	.00	.05	.01	.01	.08	.02	-.01	.05	.01	.28	(.75)					
17. Agreeableness	5.08	.83	.08	-.05	-.01	.01	.02	.03	.07	-.01	.07	-.02	.05	.05	.07	-.01	.27	.20	(.63)				
18. Conscientiousness	5.60	.72	.17	-.06	.16	.17	.15	.21	.16	.18	.20	.02	.03	.11	.07	.03	.66	.23	.33	(.61)			
19. Emotional Stability	4.82	.87	-.17	-.05	.06	.01	.07	.10	.07	.07	.08	.02	.07	.09	.10	.05	.34	.21	.38	.30	(.59)		

Table 3 continues

	Mean	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
20. Openness	4.71	.86	.07	-.08	.09	.09	.08	.15	.09	.10	.14	.02	.04	-.01	-.04	-.01	.34	.36	.34	.38	.14	(.57)	
21. Hireability	3.91	.63	.04	-.04	.15	.14	.12	.21	.13	.15	.22	.02	.02	.09	.09	.01	.76	.37	.52	.75	.46	.43	(.75)

Note: N = 733. SAT verbal and math, N = 442. ACT N = 431. College GPA N = 729. When N = 733, $p < .01$ when $r > .09$, and $p < .05$ when $r > .07$.

The GMA and verbal ability test scores generally exhibited expected patterns of intercorrelations with self-reported test scores and academic performance. Both GMA and verbal ability test scores were significantly correlated with college GPA (*r*s = .13 & .16, respectively). GMA test scores exhibited a stronger correlation with SAT math than SAT verbal test scores (*r*s = .36 & .31, respectively). Verbal ability test scores exhibited a stronger correlation with SAT verbal than SAT math test scores (*r*s = .47 & .31, respectively). Both GMA and verbal ability test scores also exhibited moderate to large correlations with ACT scores (*r*s = .42 & .52, respectively). The GMA and verbal ability test scores were moderately correlated (*r* = .43).

In Sample 2 (Table 4), gender differences were much larger for most ability variables than in Sample 1. In Sample 2, women self-reported lower IQ and standardized test scores than men, and they scored lower on the GMA and verbal ability tests than men. On the other hand, in Sample 1, only small differences were observed across genders, with women self-reporting slightly lower IQs, SAT math, and ACT scores than men. In Sample 2, self-reported IQ was much more highly related with SAT math, SAT verbal, and ACT scores (*r*s ranging from .42 to .52 at time 1 and from .38 to .59 at time 2) compared to Sample 1, perhaps due to the recency of having taken the tests for participants in Sample 2. Self-reported IQ also tended to be slightly more related to GMA and verbal ability test scores (*r*s ranging from .30 to .41 for time 1 and from .35 to .42 for time 2). Again, the GMA and verbal ability tests generally exhibited expected intercorrelations, as GMA test scores were more strongly correlated with SAT math than SAT verbal scores (*r*s .50 & .35 at time 1 and .53 & .35 at time 2, respectively), while verbal ability test scores were more strongly correlated with SAT verbal than SAT math scores (*r*s .59 & .28 at time 1 and .54 & .29 at time 2, respectively). Both GMA and verbal ability test scores were correlated with self-reported ACT (*r*s .48 & .49 at time 1, and .38 & .46 at time 2, respectively). The intercorrelations among the GMA and verbal ability tests ranged from .35 (Time 1 GMA and time 1 verbal ability) to .48 (Time 2 GMA and time 2 verbal ability).

In all, the GMA and verbal ability test scores, as well as the interviewer ratings of intellect, exhibit acceptable evidence of construct validity, with GMA test scores exhibiting the weakest evidence of reliability and validity.

Table 4. Correlation Matrix of Observed Variables in Sample 2

	Mean	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Gender	.60	.49											
2. Age	18.79	1.24	-.13										
3. Self-reported IQ: Time 1	109.93	9.88	-.26	.11									
4. Self-reported IQ: Time 2	107.91	11.14	-.30	.06	.81								
5. SAT verbal	628.33	69.47	-.19	.06	.42	.38							
6. SAT math	646.52	89.28	-.35	.04	.44	.42	.60						
7. ACT	27.75	4.27	-.27	.22	.52	.59	.62	.72					
8. GMA: Time 1	7.74	2.86	-.13	.15	.30	.36	.35	.50	.48	(.60)			
9. GMA: Time 2	8.63	3.51	-.24	.16	.30	.35	.35	.53	.38	.66	(.76)		
10. Verbal ability: Time 1	15.66	6.55	-.17	.17	.41	.42	.59	.28	.49	.35	0.39	(.79)	
11. Verbal ability: Time 2	15.72	7.02	-.09	.12	.38	.40	.54	.29	.46	.37	0.48	0.80	(.81)

Table 5. Summary of Research Questions and Results

Research Question	How	Key Results
1a: Inter-algorithm reliability (Sample 1)	r_{aa}, the correlation between predictions from Elastic net and random forest models.	r_{aa} lowest for nonverbal and paraverbal models and when predicting self-reported IQ; highest for LIWC Plus models and when predicting intellect.
1b: Split-half (across interview questions) reliability (Sample 2)	$r_{xx'}$, the correlation between predictions made separately on behavior from the odd and even numbered interview questions.	$r_{xx'}$ lowest for topic and *n*-gram models and when predicting GMA; highest for LIWC Plus and combination models and when predicting intellect.
1c: Test-retest reliability (Sample 2)	r_{tt}, the correlation between predictions made separately on behavior on Time 1 and Time 2	r_{tt} lowest for *n*-gram models and when predicting GMA; highest for LIWC Plus models and when predicting verbal ability and intellect.
2a: Convergence with ground truth	correlate ability predictions with the measure the ML model was trained to predict	*Sample 1*: Convergence was lowest for nonverbal and paraverbal models and when predicting self-reported IQ; highest for the combination models and when predicting intellect. *Sample 2*: Convergence for GMA models decreased compared to in Sample 1, but convergence for verbal ability models was relatively stable except for the topics models.
2b: Converge with ground truth stronger	Compare correlation in RQ 2a with correlations with other ability measures (i.e., GMA test, verbal ability test or Intellect)	*Sample 1*: For GMA, only the *n*-gram and combination models converged more highly with GMA than verbal ability; for verbal ability, the topics, *n*-gram, and combination models converged more highly with verbal ability than other ability measures. For intellect, all verbal behavior models converged more highly with intellect than other ability measures. *Sample 2*: For GMA, only the LIWC Plus model converged more with GMA than verbal ability; for verbal ability, all models converged more highly with verbal ability than GMA.
2c: Convergence with common proxies for	Correlate ability predictions with SAT scores, ACT scores, self-reported IQ, and college GPA	*Sample 1*: GMA and verbal ability predictions exhibited attenuated correlations with proxies compared to the test scores, whereas intellect predictions exhibited higher convergence than interviewer-rated intellect. The combination models converged highest with these proxies.

Table 5 continues

Research Question	How	Key Results
		Sample 2: Convergence with proxies was again attenuated for GMA and verbal ability models compared to test scores, but less so than in Sample 1. Intellect predictions converged more with proxies than did GMA or verbal ability predictions.
3a: Discriminant *r*s with other AVI assessments (Sample 1)	Correlate ability predictions with ability and personality predictions, compare to convergent correlations (RQ 2a)	DistilBERT and LIWC Plus models exhibited highest *r*s between ability predictions; *n*-gram and combination model exhibited the lowest. DistilBERT models tended to have the highest discriminant correlations between ability and personality predictions; *n*-gram and the combination model tended to have the lowest discriminant correlations.
3b: Discriminant *r*s with self-reported personality (Sample 1)	Correlate ability predictions with self-reports and compare to convergent correlations (RQ 2a)	The combination models' ability predictions correlated lowly with self-reported personality ($\lvert r\rvert_{max}$ = .06) and significantly with self-reported IQ (GMA r = .19; Verbal ability r = .21; intellect r = .24). Each model's convergent correlations exceeded these.
3c: Discriminant *r*s with interviewer-rated personality (Sample 1)	Correlate ability predictions with interviewer-ratings and compare to convergent correlations (RQ 2a)	The combination models' ability predictions correlated most highly with intellect ratings, then hireability, then conscientiousness, then openness, and relatively lowly with Big Five ratings. Verbal ability and intellect models' convergent correlations exceeded these discriminant ones.
4a: Group differences (Sample 1)	Compare effect size of race and gender mean differences in ground truth to ability predictions	Group differences were exacerbated in the predicted values that disadvantaged Black and East Asian interviewees but advantaged Indian ones.
4b: Accuracy across groups (Sample 1)	Compare the convergent correlations between ground truth and ability predictions across race and gender	Although no convergent correlations were significantly different, the combination models were most accurate at assessing Black interviewees. The GMA model was least accurate for Indian interviewees. The verbal ability model was least accurate for Whites. The intellect model was least accurate for women.
5a: How does ability relate to behavior (Sample 1)	Examine correlations between ability test scores and behavioral cues	GMA and verbal ability related primarily to verbal behavior, leading to more complex speech (e.g., longer words; more words; more analytical language; diverse vocabulary; more quantifiers) and appropriate responses.

Table 5 continues

Research Question	How	Key Results
5b: What cues do interviewers use to judge Ability? (Sample 1)	Visualize cue ecology for ability test scores and cue utilization for interviewer-rated intellect	Interviewers failed to utilize several specific words and phrases related to GMA and verbal ability, and their ratings were highly correlated with one invalid paraverbal behavior.
5c: What cues do algorithms use to judge ability? (Sample 1)	Visualize cue ecology for ability test scores and cue utilization for corresponding ML models	The GMA and verbal ability models rarely failed to use valid cues, yet a variety of invalid words and phrases were correlated with the ML model predictions. Additionally, even though the combination model did *not* utilize paraverbal cues, its predictions were highly correlated with one invalid paraverbal behavior.
5d: Who uses cues more validly? (Sample 1)	Compare G of ML models to G of interviewer-rated intellect	GMA and verbal ability models used cues more validity than did interviewers.
5e: Who uses cues more consistently? (Sample 1)	Compare R_s of ML models to R_s of interviewer-rated intellect	GMA and verbal ability models were more consistent in cue utilization than were interviewers.

Nested Cross-Validation (Sample 1)

The first step of the investigation involved conducting nested cross-validation and evaluating the inter-algorithm reliability and convergent evidence of the validity of algorithmic ability assessments in Sample 1. This involved creating a variety of predictive models using both elastic net regression and random forest, a variety of inputs (i.e., LIWC with readability and lexical diversity indices, referred to as LIWC Plus; LDA topics; *n*-grams; DistilBERT; nonverbal behaviors; paraverbal behaviors; and a combination of LIWC Plus and *n*-grams—the *combination model*), and using them to model GMA test scores, verbal ability test scores, self-reported IQ, and interviewer-rated intellect. Doing so provides information about RQ 1a (inter-algorithm reliability) and RQ2a (convergence with the measure of ability the algorithm was trained to model), as well as information about a) which modalities are most informative and b) which measures of ability relate most strongly to interview performance. These results are presented in Table 6.

Inter-algorithm reliability. Inter-algorithm reliability regards the convergence between the predictions from two separate ML models. In this case, one model was trained using elastic net regression, and the other was trained with random forest. In terms of modalities, inter-algorithm reliability was highest on average for LIWC Plus and DistilBERT models ($\bar{r}_{aa}$ = .88 & .87, respectively) and lowest for nonverbal and paraverbal models ($\bar{r}_{aa}$ = .63 & .69, respectively). In terms of outcomes, inter-algorithm reliability was highest on average for interviewer ratings of intellect and lowest for self-rated IQ ($\bar{r}_{aa}$ = .86 & .68, respectively).

Convergence with the ground truth. Research Question 2a regards convergence with the measure the algorithm was trained to model, or internal convergence, and represents the most basic form of convergent evidence in supervised ML. Table 6 reports the mean convergent correlations across the 10 folds, and Table 7 additionally reports the standard deviation, minimum, and maximum of the convergent correlations. Across modalities and outcomes, elastic net and random forest had, on average, approximately equal convergence ($\bar{r}$s = .286 & .283, respectively). The remaining analyses, therefore, focus on elastic net due to its lower computation time and interpretability.

Table 6. Nested Cross-Validation Convergent Correlations and Inter-Algorithm Reliability of Automated Ability Assessments by Modality and Algorithm

	GMA	Verbal Ability	Self-Report	Intellect
LIWC Plus				
r_{aa}	.89	.89	**.**85	.90
Elastic Net	.23	.37	**.24**	.52
Random Forest	.21	.35	.21	.52
Topics				
r_{aa}	.76	.87	.74	.93
Elastic Net	.25	.39	.20	.52
Random Forest	.25	.39	.16	.49
n-Grams				
r_{aa}	.75	.74	.51	.83
Elastic Net	.34	.39	.13	.46
Random Forest	.28	.40	.19	.50
DistilBERT				
r_{aa}	.84	.88	.82	.95
Elastic Net	.21	.38	.15	**.57**
Random Forest	.22	.33	.15	.54
Nonverbals				
r_{aa}	.50	.61	.65	.77
Elastic Net	.08	.12	.12	.20
Random Forest	.06	.06	.09	.21
Paraverbals				
r_{aa}	.55	.69	.71	.81
Elastic Net	.07	.18	.14	.35
Random Forest	.17	.19	.08	.37
LIWC Plus & *n*-Grams				
r_{aa}	.72	.80	.52	.85
Elastic Net	**.35**	**.40**	.15	.49
Random Forest	.29	**.40**	.23	.54

Note. $p < .01$ when $r > .09$, and $p < .05$ when $r > .07$.

Table 7. Convergent Correlations of Nested Cross-Validation Results Predicting Tested Ability for High Performing Modalities (Elastic Net)

	Convergent Correlations										
	With the Same Measure			With Other Measures (Means)							
	Mean (*SD*)	Min	Max	GMA	Verbal Ability	Self-Rated IQ	Intellect	SAT Math	SAT Verbal	ACT	College GPA
Observed											
GMA					.43	.30	.24	.36	.32	.41	.12
Verbal ability				.43		.30	.32	.29	.48	.54	.18
Intellect				.24	.32	.19		.20	.16	.30	.13
LIWC Plus											
GMA	.23 (.12)	.08	.47		.29	.17	.41	.20	.17	.31	.06
Verbal ability	.37 (.14)	.10	.60	.22		.22	.48	.19	.23	.33	.13
Intellect	.52 (.08)	.41	.65	.25	.37	.25		.19	.20	.33	.17
Topics											
GMA	.24 (.11)	.03	.37		.30	.19	.37	.22	.14	.36	.18
Verbal ability	.39 (.10)	.18	.54	.21		.21	.38	.17	.21	.40	.13
Intellect	.52 (.06)	.45	.62	.24	.33	.22		.20	.19	.40	.19
n-Grams											
GMA	.34 (.12)	.13	.51		.31	.19	.34	.22	.20	.35	.11
Verbal ability	.39 (.08)	.28	.49	.25		.20	.36	.20	.24	.36	.12
Intellect	.46 (.05)	.40	.58	.23	.29	.22		.21	.20	.35	.16

Table 7 continues

	Convergent Correlations										
	With the Same Measure			With Other Measures (Means)							
	Mean (*SD*)	Min	Max	GMA	Verbal Ability	Self-Rated IQ	Intellect	SAT Math	SAT Verbal	ACT	College GPA
DistilBERT											
GMA	.21 (.11)	.10	.38		.26	.18	.43	.13	.16	.26	.11
Verbal ability	.38 (.14)	.12	.57	.23		.19	.48	.15	.18	.32	.11
Intellect	.57 (.06)	.46	.63	.26	.34	.22		.20	.21	.34	.14
LIWC Plus & *n*-Grams											
GMA	.35 (.11)	.14	.48		.32	.19	.35	.23	.20	.36	.10
Verbal ability	.40 (.08)	.29	.52	.25		.21	.38	.21	.24	.37	.12
Intellect	.49 (.04)	.44	.59	.24	.31	.24		.22	.21	.37	.17

Note. $p < .01$ when $r > .09$, and $p < .05$ when $r > .07$.

In terms of modalities, convergence was, on average, lowest for nonverbal and paraverbal behavior models ($\bar{r}$s = .13 & .19, respectively) and highest for the combination models ($\bar{r}$ = .35). However, LIWC, topic, *n*-gram, and DistilBERT models each had similar levels of convergence ($\bar{r}$s = .34, .34, .33, & .33, respectively). For GMA and verbal ability, convergence was highest for the combination model (r = .35 & .40, respectively), although *n*-gram models had similar convergence with GMA (r = .34), and the LIWC Plus, topic, *n*-gram, and DistilBERT models each had similar levels of convergence with verbal ability (rs = .37, .39, .39, & .38, respectively). The highest convergence with GMA was lower than the accuracy of ability predictions in Kosinski et al. (2013), while the highest convergence with verbal ability was .01 higher.

Self-reports of IQ were least accurately modeled, suggesting that they were least related to interview performance. LIWC Plus models predicted self-reported IQ most accurately (r = .24). Interviewer-rated intellect was most accurately modeled, suggesting it was most related to interview performance. DistilBERT modeled interviewer-rated intellect most accurately (r = .57), although the remaining verbal behavior modalities were near .50 (ranging from .46 for *n*-gram models to .52 for LIWC Plus and topics). The convergence of interviewer-rated intellect compares favorably with the convergence observed in prior studies of AVI personality assessments (e.g., Hickman, Bosch, et al., 2021), where only conscientiousness and extraversion model predictions ever exhibited convergence exceeding .50. Considering the low convergence of self-reported IQ, nonverbal, and paraverbal behavior models, they are not further analyzed (although the nonverbal and paraverbal behaviors are still analyzed in the lens model).

Convergence with similar measures. Research question 2b regards whether convergence is higher with the measure the model was designed to assess than with other measures of ability. Table 7 summarizes these results for the accurate modalities (i.e., excluding nonverbal and paraverbal behaviors) in the first four columns under the subheading with other measures (means). For GMA, only the n-gram and combination model predictions converged more highly with GMA than verbal ability test scores ($r_{GMA} - r_{verbal}$ = .03 in both cases). The remaining three modalities' GMA predictions converged more strongly with verbal ability than GMA test scores ($r_{GMA} - r_{verbal}$ = -.05 or -.06. The GMA predictions never converged more strongly with GMA test scores than

with interviewer-rated intellect, but the n-gram and combination model GMA predictions converged as highly with GMA test scores as interviewer-rated intellect (rs = .34 & .35, respectively).

For verbal ability, every modality's predictions converged more highly with verbal ability than with GMA test scores (r_{verbal} - r_{GMA} = minimum .14 [*n*-grams], maximum .18 [topics]). However, only the topics, *n*-grams, and the combination model predictions converged more highly with verbal ability than with interviewer-rated intellect ($r_{verbal} - r_{intellect}$ = .01, .03, & .02, respectively).

For interviewer-rated intellect, every modality's predictions converged more highly with interviewer-rated intellect than GMA ($r_{intellect} - r_{GMA}$ range: .23 [*n*-grams] to .31 [DistilBERT]) or verbal ability ($r_{intellect} - r_{verbal}$ range: .15 [LIWC Plus] to .23 [DistilBERT]) test scores. Together, this evidence suggests that the GMA predictions have the weakest construct convergence and discrimination evidence relative to the verbal ability and intellect predictions.

Research question 2c regards the convergence between the ML model predictions and proxies for ability, or external convergence. In particular, this convergence can be compared to the convergence of the observed variables. Table 7 reports this information in the final four columns under the *with other measures (means)* subheading—first for the observed variables, then for the five sets of predictions.

In nearly all cases, the GMA and verbal ability predictions exhibited decreased correlations with self-reported intelligence, SAT math and verbal scores, ACT scores, and college GPA compared to the GMA and verbal ability test scores. The one exception is that topics GMA models were more strongly correlated with college GPA than GMA test scores. DistilBERT models exhibited the largest attenuation. On average, DistilBERT GMA model predictions exhibited correlations with these variables that were .134 lower than GMA test scores, and DistilBERT verbal ability model predictions exhibited correlations that were .168 lower than for verbal ability test scores. The combination model predictions exhibited the least attenuation—the GMA correlations dropped, on average, .086 and the verbal ability correlations dropped, on average, .128.

On the other hand, on average, the intellect predictions exhibited *increased* correlations with self-reported intelligence, SAT math and verbal scores, ACT scores, and college GPA. On average, the combination intellect model predictions exhibited correlations that were .046 higher

than the observed interviewer-rated intellect scores. All other modalities exhibited increases of .03 or .04 (topics).

Overall, the convergence between the predicted values and self-reported intelligence, SAT math and verbal scores, ACT scores, and college GPA was highly similar across the three variables. For example, in the combination models, GMA predictions converged, on average, r = .22 with those proxies; verbal ability predictions converged, on average, r = .23 with those proxies; and intellect predictions converged, on average, r = .24 with those proxies.

Construct discrimination and MTMM analysis. Research question 3a regards whether AVI ability assessments exhibit discriminant evidence of validity when correlated with other AVI assessments—both for ability and personality. In other words, do AVI ability measures correlate lowly with measures of other constructs. As summarized in Table 7, GMA and verbal ability test scores correlated r = .43, while GMA test scores correlated with interviewer-rated intellect r = .24, and verbal ability scores correlated with interviewer-rated intellect r = .32. These three correlations averaged r = .33. For the predicted values, the average discriminant correlations among the three predictions were highest for DistilBERT ($\bar{\boldsymbol{r}}$ = .75) and LIWC Plus models ($\bar{\boldsymbol{r}}$ = .74), while the average discriminant correlations were lowest for n-grams ($\bar{\boldsymbol{r}}$ = .57) and the combination model ($\bar{\boldsymbol{r}}$ = .59). In other words, all ability predictions are intercorrelated more than the observed variables, but predictions from some modalities are more intercorrelated than others.

Another consideration is whether the ability predictions exhibit discriminant relations with the personality predictions. Table 8 summarizes this evidence using Woehr et al.'s (2012) MTMM indices. The indices were calculated by treating each ability assessment separately, isolating only correlations relevant to the focal model's characteristics. To do so, C1 was calculated by correlating ability predictions with the measure the model was trained to predict. D1 was calculated by subtracting the heterotrait-heteromethod correlations (i.e., the average correlation between ability predictions and interviewer ratings of the Big Five traits) from C1. D2 was calculated by subtracting the average heterotrait-monomethod correlations (i.e., the average correlation between ability predictions and Big Five predictions, as well as between ability scores and Big Five ratings) from C1. Following Hickman, Bosch, et al. (2021), $D2_a$ was calculated by subtracting only the predicted heterotrait-monomethod correlations from C1. MV was calculated by subtracting D1's

heterotrait-heteromethod correlations from D2's heterotrait-monomethod correlations. MV_a was calculated using only the predicted heterotrait-monomethod correlations from $D2_a$.

Table 8. Multitrait-Multimethod Statistics by Ability for High Performing Modalities

	C1	D1	D2	$D2_a$	MV	MV_a
LIWC Plus						
GMA	.23	.04	-.01	-.19	.05	.23
Verbal ability	.37	.12	.05	-.16	.07	.28
Intellect	.52	.26	.03	-.09	.24	.35
Topics						
GMA	.25	.08	.02	-.14	.06	.22
Verbal ability	.39	.18	.09	-.11	.09	.29
Intellect	.52	.25	-.01	-.15	.26	.40
n-Grams						
GMA	.34	.20	.19	.11	.01	.08
Verbal ability	.39	.22	.16	.03	.06	.18
Intellect	.46	.19	-.02	-.12	.21	.31
DistilBERT						
GMA	.21	-.02	-.12	-.38	.09	.35
Verbal ability	.38	.13	.04	-.19	.09	.32
Intellect	.57	.28	-.11	-.11	.24	.38
LIWC Plus & *n*-Grams						
GMA	.35	.20	.19	.10	.01	.10
Verbal ability	.40	.22	.16	.03	.06	.19
Intellect	.49	.21	.01	-.10	.21	.31

Note. Calculated separately for each type of ability measure (i.e., GMA, verbal ability, and interviewer-rated intellect). For construct discrimination, interviewer-reported Big Five traits and automatically scored Big Five traits are used. C1 = convergence between predicted and observed values for the focal variable. D1 = C1 – the average correlation between predicted ability and interviewer-rated personality traits. D2 = C1 – the average correlation between (predicted ability and predicted personality as well as observed ability scores and interviewer-rated personality traits). $D2_a$ = C1 – the average correlation between predicted ability and predicted personality. MV = the second component of D2 minus the second component of D1. MV_a = the second component of $D2_a$ minus the second component of D1.

C1 was highest, on average, for the combination model ($\overline{C1}$ = .41), although DistilBERT exhibited the highest convergence for interviewer-ratings of intellect (C1 = .57). Regarding GMA models, D1 averaged .10 and ranged from -.02 (DistilBERT) to .20 (*n*-grams and the combination model). D2 average .05 and ranged from -.12 (DistilBERT) to .19 (the combination model). D2a averaged -.10 and ranged from -.38 (DistilBERT) to .10 (*n*-grams and the combination model). MV averaged .04 and ranged from .01 (*n*-grams and the combination model) to .09 (DistilBERT). MV_a averaged .20 and ranged from .08 (*n*-grams) to .35 (DistilBERT).

For verbal ability models, D1 averaged .17 and ranged from .12 (LIWC Plus) to .22 (*n*-grams and the combination model). D2 averaged .10 and ranged from .04 (DistilBERT) to .17 (*n*-grams). $D2_a$ averaged -.08 and ranged from -.19 (DistilBERT) to .03 (*n*-grams and the combination model). MV averaged .07 and ranged from .06 (*n*-grams and the combination model) to .09 (topics and DistilBERT). MV_a averaged .25 and ranged from .18 (*n*-grams) to .32 (DistilBERT).

For interviewer-rated intellect models, D1 averaged .24 and ranged from .19 (*n*-grams) to .28 (DistilBERT). D2 averaged -.02 and ranged from -.11 (DistilBERT) to .03 (LIWC Plus). $D2_a$ averaged -.11 and ranged from -.15 (topics) to -.09 (LIWC Plus). MV averaged .23 and ranged from .21 (*n*-grams and the combination model) to .26 (topics). MV_a averaged .35 and ranged from .31 (*n*-grams and the combination model) to .40 (topics). Considering that DistilBERT models exhibited such poor construct discrimination, I do not further analyze them.

Research Question 3b regards whether the algorithmic ability assessments exhibit expected discriminant relationships with self-reported Big Five traits. First, I consider the characteristics of the self-reports themselves. Self-reports of IQ were largely independent of self-reported Big Five traits (*r*s ranging from -.05 to .03). Meta-analytic results suggest that self-reported Big Five traits are largely independent of ability (e.g., Schilling et al., 2021).

Here I summarize discriminant results for the combination model, considering that it had the best convergent and discriminant evidence according to the MTMM indices. For the GMA model, the strongest correlation with self-reports was with self-reported IQ (r = .19). The correlations between GMA predictions and Big Five traits ranged from -.03 (openness) to .01 (emotional stability). Considering the verbal ability model, the strongest correlation with self-reports was with self-reported IQ (r = .21). The correlations between verbal ability predictions and Big Five traits ranged from -.03 (emotional stability) to .06 (conscientiousness). Considering the intellect model, the strongest correlation with self-reports was with self-reported IQ (r = .24). The

correlations between intellect predictions and Big Five traits ranged from -.00 (agreeableness) to .05 (emotional stability).

Research Question 3c regards whether the algorithmic ability assessments exhibit expected discriminant relationships with interviewer-rated Big Five traits and hireability. First, I consider the characteristics of the interviewer ratings themselves. Interviewer ratings of conscientiousness, intellect, and hireability were highly correlated ($\bar{r}$ = .72). The remaining correlations with ratings of hireability ranged from .37 (extraversion) to .51 (agreeableness), and the remaining intercorrelations among intellect and Big Five trait ratings averaged .29 and ranged from .13 (emotional stability – openness) to .39 (conscientiousness – openness). Therefore, we would expect the highest discriminant correlations to be with conscientiousness and hireability and for the remaining discriminant correlations to be lower. Additionally, openness is often correlated with ability because openness involves one's propensity to engage in intellectual tasks and abstract thinking.

Here I summarize such results for the combined model, as presented in Table 9. Considering first the GMA models, all correlations with interviewer ratings were lower than C1 except for intellect, where r = C1 = .35. The highest correlations with the remaining interviewer ratings were with conscientiousness (r = .29) and hireability (r = .28). The correlations with the remaining ratings averaged .12 and ranged from .08 (agreeableness) to .17 (openness).

Considering the verbal ability models, all correlations with interviewer ratings were lower than C1. The highest correlations with interviewer ratings were for intellect (r = .38), hireability (r = .34), and conscientiousness (r = .30). The correlations with the remaining ratings averaged .15 and ranged from .11 (agreeableness) to .23 (openness).

Considering the intellect models, the remaining correlations with interviewer ratings were lower than C1. The highest correlations with the remaining interviewer ratings were with hireability (r = .45) and conscientiousness (r = .42). The correlations with the remaining ratings averaged .24 and ranged from .13 (agreeableness) to .32 (openness). Overall, the models exhibited expected intercorrelations with self-reports and interviewer-ratings.

Table 9. Nested Cross-Validation Correlation Matrix for the LIWC Plus and *n*-Grams Combination Model

	1	2	3	4	5	6	7	8	10	11	12
Combination Model											
1. GMA											
2. Verbal	.61										
3. Intellect	.50	.64									
4. Extraversion	.21	.33	.59								
5. Agreeableness	.07	.26	.38	.33							
6. Conscientiousness	.44	.54	.84	.58	.43						
7. Emotional Stability	.19	.28	.50	.21	.35	.41					
8. Openness	.32	.44	.62	.70	.44	.63	.22				
Ability Scores											
9. GMA	.35	.25	.24	.08	.07	.21	.12	.12			
10. Verbal	.32	.40	.31	.13	.12	.26	.12	.21	.43		
11. Self-rated IQ	.19	.21	.24	.13	.01	.17	.08	.09	.30	.30	
12. Intellect	.35	.38	.49	.33	.16	.45	.32	.35	.24	.32	.19
Interviewer Ratings											
Extraversion	.09	.13	.28	.42	.18	.27	.15	.34	.00	.02	.14
Agreeableness	.08	.11	.13	.17	.23	.15	.18	.21	.00	.08	.00

Table 9 continues

	1	2	3	4	5	6	7	8	10	11	12
Conscientiousness	.29	.30	.42	.29	.20	.45	.23	.34	.18	.21	.17
Emotional Stability	.13	.13	.22	.12	.16	.18	.40	.10	.07	.09	.06
Openness	.17	.23	.32	.33	.22	.31	.14	.42	.10	.15	.10
Hireability	.28	.34	.45	.35	.25	.42	.32	.37	.16	.23	.16
Self-Reports											
Extraversion	-.01	.00	.04	.06	.07	.06	.06	.03	.02	-.08	-.05
Agreeableness	-.01	-.02	.00	.00	.07	.01	.06	.01	.01	-.04	.00
Conscientiousness	.00	.06	.04	.01	.02	.04	.06	.00	.07	.06	.03
Emotional Stability	.01	-.03	.05	.09	.04	.06	.09	.02	.09	-.01	.01
Openness	-.03	.05	.01	-.02	.03	.01	.04	-.02	.06	.09	.01

Note. Calculated by averaging together the correlation matrices from each test fold. Suppressed columns are in Table 3.

Group differences and bias. Table 10 reports a) the GMA test scores, verbal ability test scores, and intellect ratings broken down by race and gender (upper section), and b) the combination model's GMA, verbal ability, and intellect predictions broken down by race and gender (lower section) to address RQ 4a. Considering race/ethnicity, in the observed values, Black and African American participants scored lowest on average on all three measures, while Whites scored highest on GMA and verbal ability, and Indians were rated highest on intellect (several groups, including Native Hawaiian or Pacific Islander, American Indian or Alaska Native, Middle Eastern, and African were not included in this analysis due to small N, ranging from 2 to 16). The Black-White ds were the largest and ranged from -.42 (intellect) to -.77 (GMA and verbal ability), and the Hispanic-White differences were also sizeable, ranging from -.25 (verbal ability) to -.36 (GMA). Men scored slightly higher than women on average on GMA (d = -.07), while women scored slightly higher than men on verbal ability (d = .06) and intellect (d = .01).

In the predicted values from the ML models, Black and African American participants again scored lowest on all three measures. Whites again scored highest on verbal ability. Differing from the observed data, Indians scored highest on GMA and tied with Whites for the highest average intellect scores. Black-White differences increased slightly in the predicted values (Δds ranging from -.04 to -.06 [GMA and intellect]), as did East Asian-White differences (Δds ranging from -.03 [intellect] to -.06 [verbal ability]). On the other hand, Hispanic-White differences decreased (Δds ranging from .05 [verbal ability] to .24 [intellect]), and Indian-White differences decreased with Indians now scoring higher than Whites on GMA, d changing to -.07 for verbal ability, and Indians no longer scoring higher than Whites on intellect. Gender differences were reduced for GMA in the predicted values (Δd = .06) but increased for verbal ability (Δd = .11) and intellect (Δd = .13).

Table 11 reports the convergence between observed and predicted ability scores by race and gender to address RQ 4b. For each variable, the predictions were most accurate for Black and African American interviewees (r_{GMA} = .48; r_{verbal} = .45; $r_{intellect}$ = .55). Accuracy varied most widely for GMA, with predictions being least accurate for Indians (r = .23). Verbal ability predictions were least accurate for Whites (r = .33), and intellect predictions were least accurate for women (r = .45). However, no pairwise correlations were significantly different at $p < .05$ when using Fisher's r-to-z transformation.

Table 10. Ability Scores Analyzed by Race and Gender for the Observed Values and Combination Model

Observed Scores	East Asian Mean (*SD*)	Black Mean (*SD*)	Hispanic Mean (*SD*)	Indian Mean (*SD*)	White Mean (*SD*)	Women Mean (*SD*)	Men Mean (*SD*)
GMA	8.47 (3.26)	6.23 (2.95)	7.39 (3.25)	8.02 (3.28)	8.54 (3.06)	8.10 (3.17)	8.33 (3.34)
Verbal Ability	18.42 (7.56)	14.89 (6.93)	18.35 (8.37)	18.02 (8.15)	20.34 (7.26)	19.20 (7.54)	18.75 (7.65)
Intellect	5.49 (.61)	5.19 (.80)	5.33 (.59)	5.62 (.64)	5.50 (.67)	5.47 (.66)	5.46 (.68)
	East Asian-White Cohen's *d*	Black-White Cohen's *d*	Hispanic-White Cohen's *d*	Indian-White Cohen's *d*		Women-Men Cohen's *d*	
GMA	-.02	-.77	-.36	-.16		-.07	
Verbal Ability	-.26	-.77	-.25	-.30		.06	
Intellect	-.02	-.42	-.27	.18		.01	
Predicted Values	East Asian Mean (*SD*)	Black Mean (*SD*)	Hispanic Mean (*SD*)	Indian Mean (*SD*)	White Mean (*SD*)	Women Mean (*SD*)	Men Mean (*SD*)
GMA	8.25 (1.24)	7.16 (1.40)	8.14 (1.25)	8.47 (1.26)	8.34 (1.45)	8.19 (1.42)	8.21 (1.35)
Verbal Ability	18.66 (3.55)	17.00 (3.33)	19.06 (3.62)	19.56 (3.03)	19.79 (3.55)	19.40 (3.66)	18.82 (3.27)
Intellect	5.48 (.35)	5.32 (.36)	5.49 (.37)	5.50 (.32)	5.50 (.39)	5.49 (.37)	5.44 (.37)

Table 10 continues

	East Asian-White Cohen's *d*	Black-White Cohen's *d*	Hispanic-White Cohen's *d*	Indian-White Cohen's *d*	Women-Men Cohen's *d*
GMA	-.07	-.83	-.15	.10	-.01
Verbal Ability	-.32	-.81	-.20	-.07	.17
Intellect	-.05	-.48	-.03	.00	.14

Note. East Asian *N* = 171; Black *N* = 56; Hispanic *N* = 54; Indian *N* = 44; White *N* = 364; Women *N* = 465; Men *N* = 262.

Table 11. Correlational Accuracy of Ability Predictions by Race and Gender for the Combination Model

	GMA	Verbal Ability	Intellect
Overall	.35	.40	.49
East Asian	.26	.44	.48
Black	.48	.45	.55
Hispanic	.42	.41	.46
Indian	.23	.34	.48
White	.33	.33	.49
Women	.35	.41	.51
Men	.37	.37	.45

Note. No correlations are significantly different at $p < .05$ using Fisher's *r*-to-*z* transformation, but power is very limited for all ethnic comparisons.

Cross-Sample Cross-Validation (Sample 2)

As an additional step in investigating the psychometric properties of algorithmic ability assessments, I trained models on the full set of participants in the main sample and then applied those models to assess ability at both time points in the test-retest sample.

Split-half reliability. To address Research Question 1b, Table 12 reports the split-half reliability (r_{xx}') of the algorithmic ability assessments made at Time 1. For GMA, split-half reliability averaged .25 and ranged from .23 (n-grams) to .27 (the combination model). For verbal ability, split-half reliability averaged .35 and ranged from .26 (n-grams) to .46 (LIWC Plus). For intellect, split-half reliability averaged .47 and ranged from .22 (topics) to .58 (the combination model).

Test-retest reliability. To address Research Question 1c, Table 12 reports the test-retest reliability (r_{tt}) of the algorithmic ability assessments. For GMA, test-retest reliability averaged .40 and ranged from .27 (*n*-grams and the combination model) to .59 (topics). For verbal ability, test-retest reliability averaged .56 and ranged from .50 (*n*-grams) to .64 (LIWC Plus). For intellect, test-retest reliability averaged .55 and ranged from .44 (topics) to .65 (LIWC Plus). The values for verbal ability and intellect are similar to, but slightly smaller than, the most optimistic test-retest reliability estimates reported for AVI personality assessments by Hickman et al. (2021). In all cases, test-retest reliability was equal to or larger than split-half reliability.

Cross-sample convergent evidence. Next, I investigated the validity evidence of the ability assessments in Sample 2. To do so, I averaged together the two GMA and verbal ability test scores and predictions for GMA, verbal ability, and intellect made at Times 1 and 2. This evidence is summarized in Table 12. First, it appears that the topics models did not work correctly when moving to the test-retest sample. The topics were generated in the main sample, but the topic model does not appear to be robust and generalizable to the test-retest sample. This may be because the participants in the main sample tended to be older and later in their undergraduate studies than the participants in the test-retest sample. The age difference may have resulted in different topics being emphasized in their responses, since participants in the test-retest sample have few workplace experiences to draw on in their responses. As a result, I do not further investigate the topics models.

The test-retest sample does not have interviewer ratings, so I focus on the convergence between ability predictions, ability test scores, and similar measures. To address Research Questions 2a and 2b, I summarize the convergence between the average of the two predictions and ability test scores. For GMA, predictions converged, on average, .21 with GMA test scores and .27 with verbal ability test scores. Only the LIWC Plus model predictions converged more highly with GMA than with verbal ability, but the difference in correlation was only .01. The verbal ability predictions converged, on average, .40 with verbal ability test scores and .20 with GMA test scores. All the model predictions of verbal ability converged more highly with verbal ability than GMA, with differences ranging from .22 (the combined model) to .20 (LIWC and *n*-grams). The intellect predictions converged, on average, .23 with GMA and .36 with verbal ability. In the *n*-gram and combination models, the intellect predictions converged more highly with GMA test scores than the GMA predictions did.

Table 12. Cross-Sample Reliability, Convergent, and Discriminant Evidence of Validity (Combination Model)

Reliability		Convergence						
							Self-Reported IQ	
	$r_{xx'}$	r_{tt}	GMA	Verbal	SAT Verbal	SAT Math	ACT	Time 1	Time 2
LIWC Plus									
GMA	.25	.45	.28	.27	.30	.27	.34	.22	.21
Verbal	.46	.64	.17	.37	.29	.11	.32	.23	.24
Intellect	.56	.65	.22	.39	.32	.22	.40	.28	.25
n-Grams									
GMA	.23	.27	.17	.26	.16	.18	.35	.20	.20
Verbal	.26	.50	.21	.41	.29	.14	.28	.18	.16
Intellect	.53	.53	.23	.33	.27	.21	.37	.21	.18
Topics									
GMA	.25	.59	.08	.07	-.01	-.05	-.18	-.06	-.06
Verbal	.34	.53	.01	.03	-.04	-.01	-.12	-.06	-.09
Intellect	.22	.44	.04	.04	.00	-.05	-.17	.00	-.00
LIWC Plus & *n*-Grams									
GMA	.27	.27	.19	.27	.21	.20	.37	.22	.22
Verbal	.33	.56	.21	.43	.31	.14	.30	.19	.17
Intellect	.58	.59	.23	.35	.29	.22	.38	.22	.20

Note. Convergence calculated by averaging the time 1 and time 2 scores for GMA, Verbal ability, GMA predictions, and Verbal ability predictions, then correlating the same-trait scores. $r_{xx'}$ is split-half reliability, calculated by making predictions on the Time 1 behavior from the odd numbered questions and even numbered questions separately, then correlating those two predictions. r_{tt} is test-retest reliability, calculated by making predictions on the Time 1 and Time 2 responses separately, then correlating those two predictions.

To address Research Question 2c, I examine the convergence between the average of the two ability predictions and proxies for ability. The correlations between the GMA and verbal ability predictions and proxies for ability were again substantially attenuated compared to the observed correlations. Whereas GMA test scores and verbal ability test scores correlated at least .50 with SAT math and SAT verbal scores, respectively, the average of the two predictions converged, at most, .31 with SAT verbal (the combination verbal ability model) and .27 with SAT math (the LIWC Plus GMA model). Similarly, the observed test scores converged with ACT scores at least r = .38, yet the largest correlation between GMA or verbal ability predictions and ACT scores was r = .37 (the combination GMA model).

The intellect predictions exhibited stronger convergent correlations with SAT verbal and ACT scores than did either GMA or verbal ability predictions. The average of the LIWC Plus model's intellect scores converged r = .32 with SAT verbal, r = .40 with ACT scores, and r = .27 with self-reported IQ. In the combination models, GMA predictions converged, on average, r = .25 with these proxies. Verbal ability predictions converged, on average, r = .23 with these proxies, and the intellect predictions converged, on average, r = .28 with these proxies. The convergence with proxies is slightly higher for GMA compared to the within-sample investigations (Δr = .03), the same for verbal ability (Δr = .00), and slightly higher for intellect (Δr = .04), but any increases may be due to the enhanced measurement reliability from averaging the two AVI assessments. Convergence with both ability test scores and proxies was actually highest, on average, for the LIWC Plus models in the test-retest sample.

Brunswik Lens Model Analysis (Sample 1)

Research Question 5a regards how ability relates to interviewee behavior in Sample 1. To investigate this, I draw on the Brunswik Lens model to illustrate the relationship between key behaviors, ability, and ability predictions in Figures 5-8. Figures 5 and 7 provide information about GMA, and Figures 6 and 8 provide information about verbal ability. The figures cannot contain all behaviors investigated since the present study included over 5,000 predictor variables across all modalities. Therefore, to help address Research Question 5a, Table 13 provides the 30 strongest correlations between behavior, GMA, and verbal ability test scores in Sample 1. For the lens model equations, a subset of approximately 250 behavioral cues most strongly correlated with GMA and verbal ability respectively were used in the calculations.

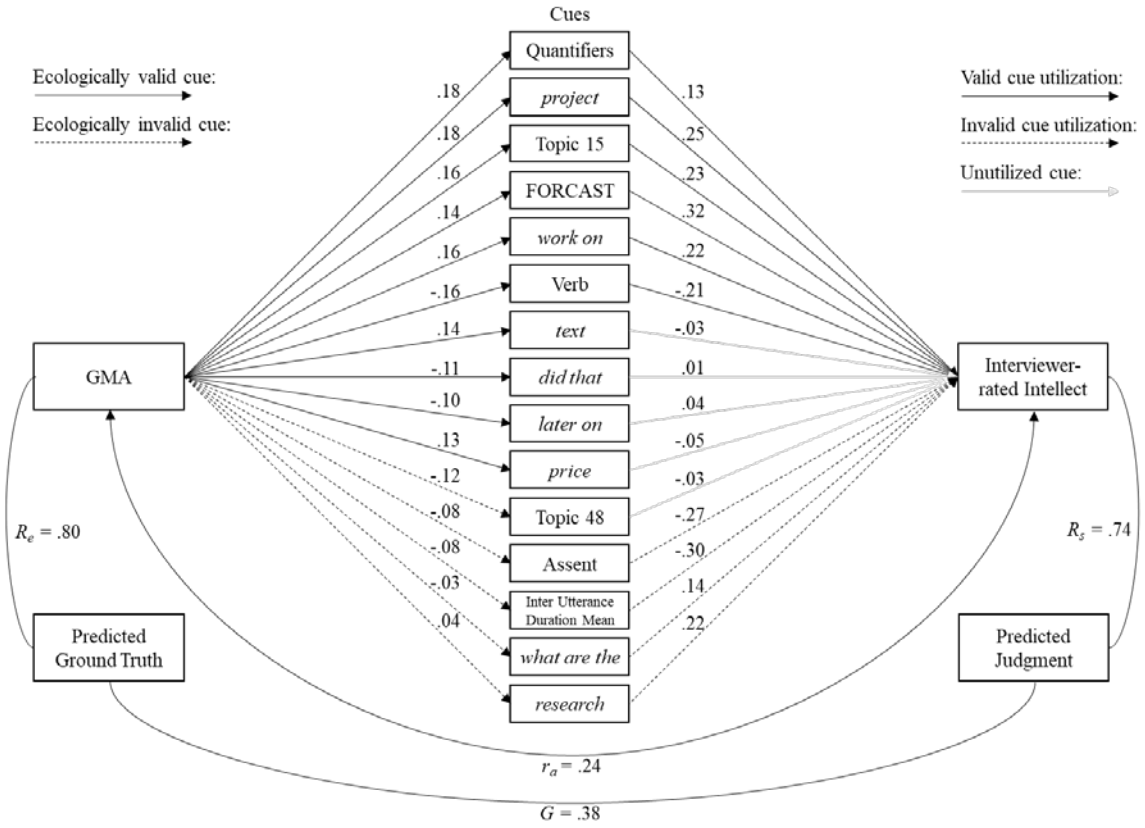


Figure 5. Brunswik lens model of GMA scores and interviewer-rated intellect.

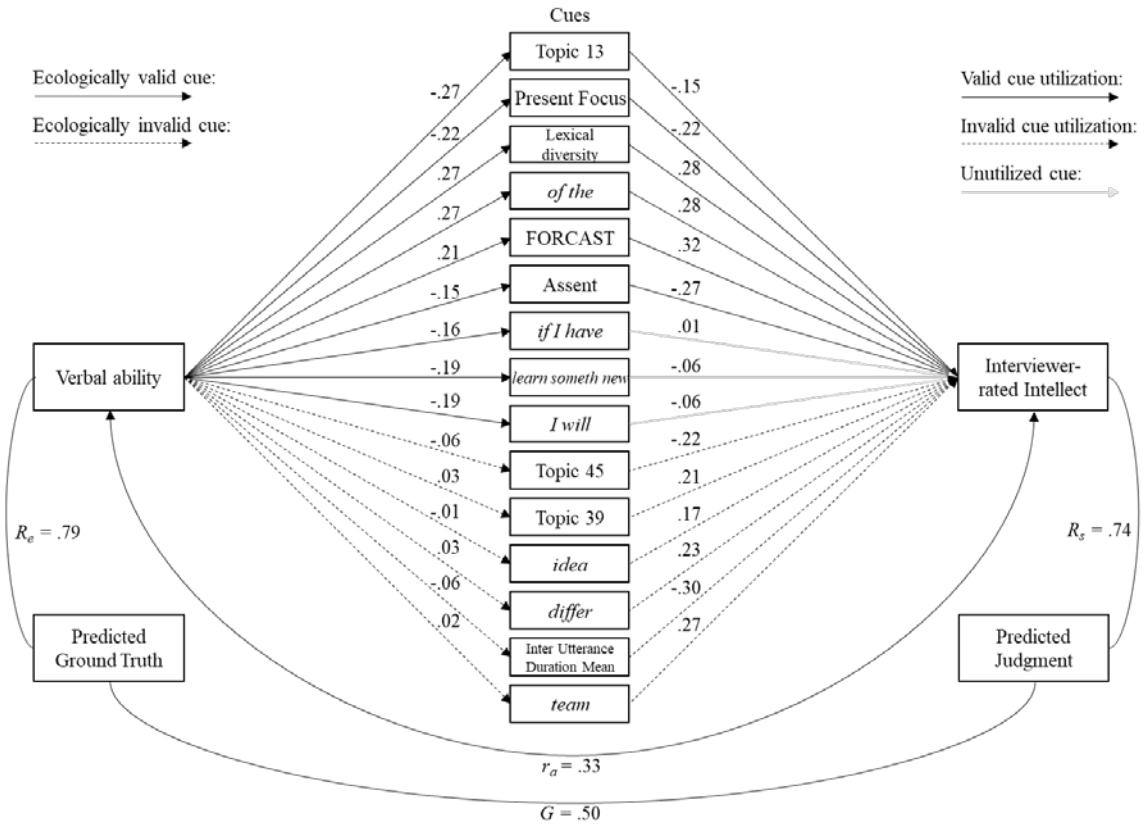


Figure 6. Brunswik lens model of verbal ability scores and interviewer-rated intellect.

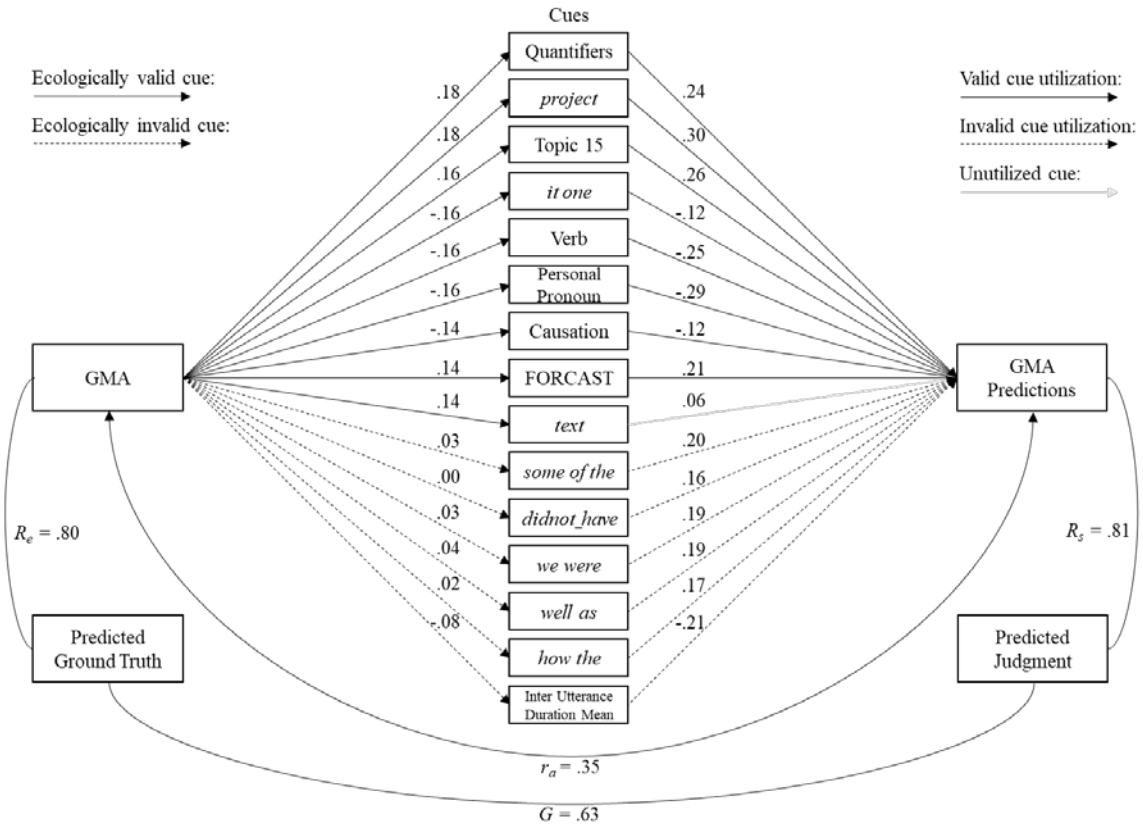


Figure 7. Brunswik lens model of GMA scores and combination model GMA predictions.

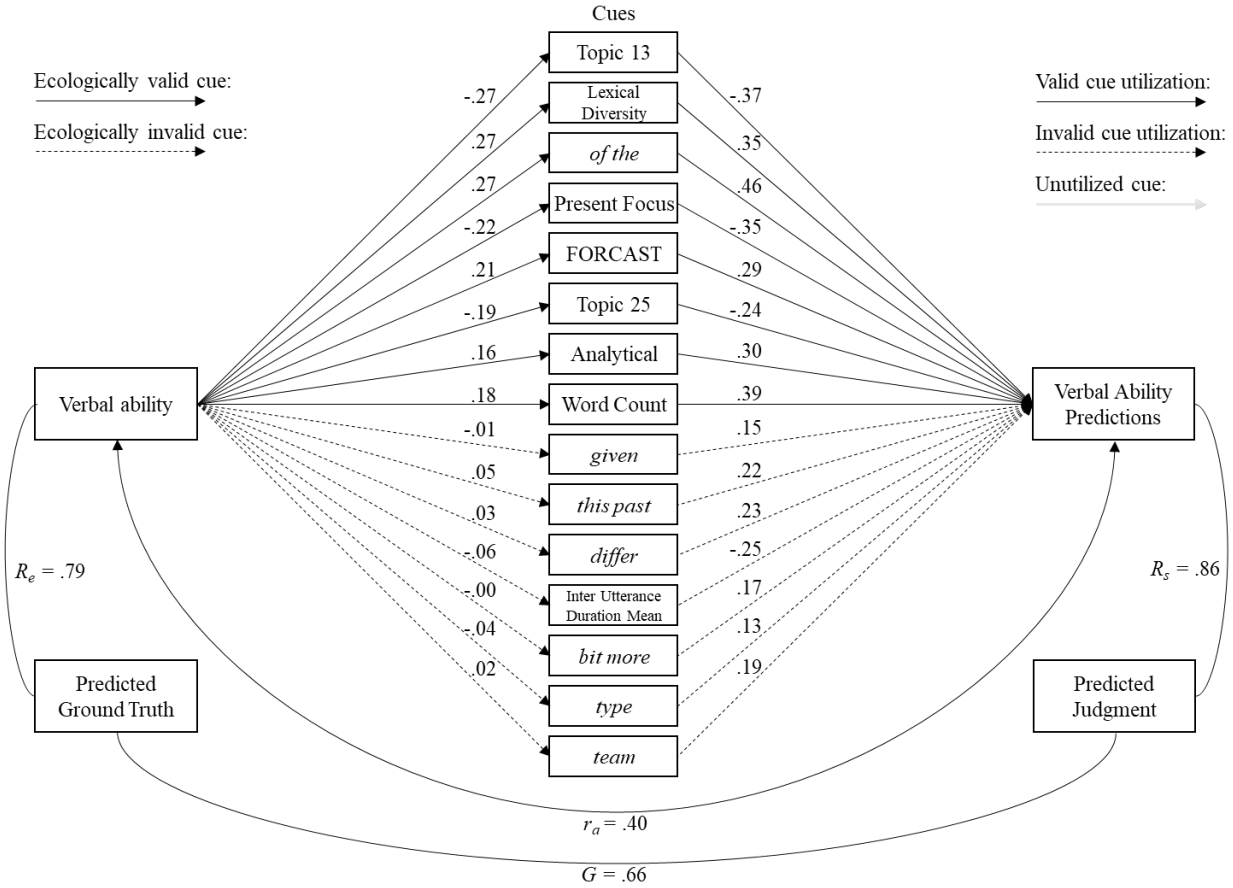


Figure 8. Brunswik lens model of verbal ability scores and combination model verbal ability predictions.

Table 13. Cues Most Strongly Related to GMA and Verbal Ability Test Scores

GMA		Verbal Ability	
Cue	*r*	Cue	*r*
Quantifiers	.18	Topic 13	-.27
project	.18	Lexical Diversity	.27
it one	-.16	*of the*	.27
Topic 15	.16	Present Focus	-.22
Common Verbs	-.16	FORCAST	.21
of the	.16	Words > 6 Letters	.20
work on	.16	Topic 25	-.19
that i could	.16	*i will*	-.19
the main	.16	Personal Pronouns	-.19
Personal Pronouns	-.16	*learn someth new*	-.19
or just	.15	Mean Syllables Per Word	.19
higher	.15	*someth new*	-.19
part of	.15	Assent SD	-.19
Auxiliary Verbs	-.15	*and so*	.18
over the	.15	*thing*	.18
we need	.15	Function Words SD	-.18
and this was	.15	Word Count	.18
specif	.15	Topic 40	.17
to make sure	.15	Personal Pronouns SD	-.17
main	.14	*one of*	.17
and a lot	.14	*consid*	.17
2nd Person Pronouns SD	-.14	Pronouns sd	-.17
project and	.14	1st Person Singular Pronouns	-.17
Causation SD	-.14	*i was*	.16
sort of	.14	Common Verbs	-.16

Table13 continues

GMA		Verbal Ability	
Cue	*r*	Cue	*r*
FORCAST	.14	Articles SD	-.16
Words > 6 Letters	.14	2nd Person Pronouns SD	-.16
2nd Person Pronouns	-.14	Analytical Thinking	.16
text	.14	*of a*	.16
i would be	.13	*if i have*	-.16

Note. *Italics* indicates a stemmed *n*-gram.

Environmental side of the lens. For GMA, 178 predictors were correlated $r > |.10|$. The strongest single correlation was with the quantifiers LIWC category (r = .18; example words include few, add, percent). The second strongest correlation was with the n-gram project (r = .18), and the fourth strongest correlation was with Topic 15 which includes the terms project, work, team, and internship (r = .16). Several LIWC categories were negatively related to GMA, including personal pronouns (r = -.16), auxiliary verbs (r = -.15), 2nd person pronouns (e.g., you, your; r = -.14), perceptual processes (r = -.11), focus present (r = -.11), pronouns (r = -.12), you (r = -.14), and the standard deviation in the use of causation words (e.g., because, effect; r = -.14), interrogatives (r = -.12), informal language (r = -.11), and power drive (r = -.11). The FORCAST readability index and the use of words longer than six letters (a LIWC category) were both positively related to GMA (rs = .14), as were average word count and analytical thinking (rs = .13). Additionally, although they fall outside the top 30 strongest correlations, facial action units 14, 17, and 28, as well as facial expressions of anger, were all negatively correlated with GMA (-.13 < rs < -.11).

Additionally, some paraverbal behaviors were related to GMA but fell outside the top 30 strongest correlations. The standard deviation of Mel-Frequency-Cepstral-Coefficients, which represents variation in volume and frequency, was correlated r = -.10 with GMA. On the other

hand, the means of jitter and shimmer, which represent variation in frequency and amplitude, respectively, were both positively correlated with GMA *r*s = .11.

For verbal ability, 344 predictors were correlated $r > |.10|$. Verbal ability was positively correlated with lexical diversity, the FORCAST readability index, the use of words longer than six letters, and the mean number of syllables in words (*r*s = .27, .21, .20, and .19, respectively). Average word count across questions was also positively correlated with verbal ability ($r = .18$). Several topics were correlated with verbal ability. Topic 13 had the strongest negative correlation with verbal ability ($r = -.27$) and included terms such as *I have, will, feel, have to, someth,* and *can*. Topic 25 was negatively correlated with verbal ability ($r = -.19$) and included terms such as *help, help them, to help, peopl, person,* and *ask*. Topic 40 was positively correlated with verbal ability ($r = .17$) and included terms such as *and so, so I, and then, that I,* and *I was*. Several LIWC categories were negatively correlated with verbal ability, including focus present (e.g., *today, is, now*, $r = -.22$), personal pronouns ($r = -.19$), verbs ($r = -.16$), auxiliary verbs ($r = -.15$), focus future ($r = .14$), pronouns ($r = -.13$), perceptual processes ($r = -.13$), and social processes ($r = -.13$). Additionally, the standard deviation of the assent (*agree, OK, yes,* $r = -.19$), function words ($r = -.18$), personal pronouns ($r = -.17$), pronouns ($r = -.17$), I ($r = -.17$), articles ($r = -.16$), you ($r = -.16$), informal language ($r = -.15$), power drives ($r = -.15$), adjectives ($r = -.14$), and negator ($r = -.13$) categories were also negatively correlated with verbal ability. Several LIWC categories were positively correlated with verbal ability, including analytical thinking ($r = .16$), prepositions ($r = .15$), and quantifiers ($r = .14$).

Few nonverbal or paraverbal behaviors were related to verbal ability. Action unit 17, which was negatively related to GMA, was also negatively related to verbal ability ($r = -.11$). The standard deviation of action unit 20, the lip stretcher, was negatively related to verbal ability ($r = -.10$). Jitter and shimmer were positively related to verbal ability (*r*s = .11 & .13, respectively). Speech rate was also positively related to verbal ability ($r = .11$). Overall, GMA and verbal behavior related primarily to differences in what was said and the characteristics of what was said, rather than how it was vocalized (i.e., paraverbal behaviors) or what the interviewee did while saying it (i.e., nonverbal behaviors).

Cue utilization. The Brunswik lens model facilitates an examination of both the match in cue ecology and cue utilization, as well as a visualization of the specific cues that were well utilized and those that were not, to address RQs 5b-5e. First, I use the Brunswik lens to explore four types of cues: 1) highly ecologically valid cues that were utilized by the rater; 2) highly ecologically valid cues that were not utilized by rater; and 3) invalid cues utilized by the rater (note, however, that invalid cues were not included when calculating the Brunswik lens equations). The intellect ratings are compared to the GMA and verbal ability environments (RQ 5b) in Figures 5 and 6 with 15 representative behaviors. These two figures begin with valid cues utilized to judge intellect, then valid cues not utilized in those judgments, and finally, invalid cues used in those judgments. For example, FORCAST was a valid cue for both GMA and verbal ability and was correlated r = .32 with intellect ratings. Several n-grams related to GMA, such as text and price, and several n-grams related to verbal ability, such as learn someth new, were not related to intellect judgments. Several invalid cues were utilized by raters as well, including inter utterance duration mean (i.e., the sum of time between utterances divided by the number of utterances, or average length of pauses), which was invalid for both GMA and verbal ability. Most other invalid cues utilized by interviewers, however, tended to be correlated at least $r = .10$ with the other ability measure. For example, although the n-gram research was only correlated $r = .04$ with GMA, it was correlated $r = .12$ with verbal ability.

Figures 7 and 8 present the environmental and judgment sides of the lens for GMA and GMA predictions, and verbal ability and verbal ability predictions, respectively (RQ 5c). Overall, predicted values tended to utilize a greater proportion of the available valid cues than did interviewers. For example, the GMA predictions failed to use very few valid cues, with the *n*-gram *text* being one example. However, the GMA predictions also tended to be correlated with several invalid *n*-grams, such as *some of the, we were,* and *well as*. Additionally, although GMA predictions from the combined model did *not* include it as a predictor, inter utterance duration mean was negatively correlated with the predictions.

Verbal ability predictions tended to be correlated with all the valid cues in the expected direction, although generally with stronger correlations than in the environmental side of the lens. For example, word count was more than twice as strongly related to verbal ability predictions than to verbal ability test scores. Verbal ability predictions were also correlated with several invalid

n-grams, including *team*, *type*, and *differ*. Additionally, although verbal ability predictions did not include it as a predictor, they were also negatively correlated with inter utterance duration mean.

Lens model equations. Second, Table 14 provides the Brunswik lens statistics. G is the correlation between the predictions of the linear models of cue ecology and cue utilization and speaks to the validity of cue utilization (RQ 5d). *G* was highest for the two sets of ML models and lowest for the GMA – Intellect comparison. Because approximately 250 predictors were empirically selected for analysis based on their correlations with the respective test scores, $\boldsymbol{R_e}$ (the upper limit of the environmental predictability), was very similar for GMA and verbal ability (i.e., .79 & .80, respectively). $\boldsymbol{R_s}$ regards how consistently cues are utilized to judge interviewees (RQ 5e), and both ML models were more consistent than the interviewers, with the verbal ability model being the most consistent. The $\boldsymbol{R_s}$ value for intellect ratings suggests the upper limit on their predictability is .74. $\boldsymbol{C}$ is non-zero when cues used by the judges were not included in the linear models and/or when cues irrelevant to the observer's judgment were included in the linear models. Hence, some cues used by interviewers were not included in the linear models, and cues not used by the ML models, including nonverbals, paraverbals, and topics, were included in the linear models. Since the composite indices, $\boldsymbol{GR_e}$ and $\boldsymbol{GR_s}$, are partially a function of knowledge, they tended to be higher for the ML algorithms. The $\boldsymbol{GR_e}$ values suggest that more consistent judges could increase the convergence between intellect and GMA to .30 and between intellect and verbal ability to .40, while GMA predictions could converge as highly as .50 and verbal ability predictions as highly as .52.

Table 14. Brunswik Lens Model Analyses

	G	R_e	R_s	C	r_a	GR_e	GR_s
GMA – Interviewer-rated Intellect	.38	.80	.74	.03	.24	.30	.28
Verbal – Interviewer-rated Intellect	.50	.79	.74	.08	.33	.40	.37
GMA – GMA predictions	.63	.80	.81	-.16	.35	.50	.51
Verbal ability – Verbal ability predictions	.66	.79	.86	-.16	.40	.52	.57

Note. Using the 251 and 249 predictors most strongly correlated with GMA and Verbal ability, respectively. G = matching index, or knowledge. R_e = the upper limit of environmental predictability. R_s = the consistency with which judges execute decision rules. C = the correlation between the residuals of the two models. r_a = achievement, or the correlation between observed and judged values. GR_e = the validity of a model created by replacing a judge with their strategy. GR_s = performance, or the rater's contribution to r_a.

DISCUSSION

The present study took several steps to advance our understanding of both AVIs and employment interviews. Prior studies of AVIs have focused on noncognitive predictors of job performance (e.g., personality). Moreover, research validating AVIs is still in its nascent stages and mainly focuses on convergent evidence of validity rather than holistically examining AVIs' psychometric properties (e.g., reliability, discriminant-related validity evidence, bias, and content). This study aimed to investigate the reliability and validity of automated ability assessments, and to advance our understanding of how interview performance relates to both ability and ability judgments. To the extent that AVIs for assessing ability can be paired with AVI personality assessments, it may expand the utility of AVIs by enabling them to serve as a one-stop shop for assessing a range of important KSAOs.

First, ML models were trained and tested using nested cross-validation to predict four operationalizations of ability using six different sets of predictors and a combination of two of the sets to investigate inter-algorithm reliability, convergent evidence of validity, discriminant evidence of validity, and potential bias. Inter-algorithm reliability was highest for the LIWC Plus models and lowest for the nonverbal behavior models, and in terms of measures, was highest for intellect and lowest for self-reported IQ. Convergence between ML model predictions and ground truth was strongest for the verbal behavior modalities, and predictions of intellect ratings were more accurate than predictions of GMA or verbal ability test scores. Convergence with the ground truth was at least as high as convergence with other measures of ability only for the *n*-gram and combination models. Regarding proxies for ability, the GMA and verbal ability predictions exhibited substantially attenuated correlations (i.e., standardized test scores and college GPA) compared to the test scores themselves. However, intellect predictions exhibited higher correlations with these proxies than did the ratings themselves. In both cases, convergence with proxies was highest for the combination model. The MTMM indices for construct discrimination and method variance were worst for DistilBERT models and best for the combination models, on average. Overall, ability predictions from the combination model exhibited expected intercorrelations with self-reports and interviewer-ratings. The combination model was most accurate as assessing Black interviewees, yet the lower accuracy in assessing some other groups

(e.g., Indians) may have disproportionately increased their scores relative to Black interviewees, thereby disadvantaging Black interviewees.

Second, ML models were trained on the entirety of Sample 1 and used to assess participants in Sample 2. Split-half reliability was highest for LIWC Plus models, as expected, but was quite low for GMA. Similarly, test-retest reliability was highest for LIWC Plus models but was quite low for GMA for some modalities. When taking the average of the predicted scores from Times 1 and 2 in Sample 2, GMA predictions did not converge more highly with GMA than verbal ability test scores, except for the LIWC Plus GMA model. GMA and verbal ability predictions again exhibited attenuated correlations with standardized test scores compared to the actual test scores. Intellect predictions converged more strongly with SAT and ACT scores than did either GMA or verbal ability predictions.

Third, I used the Brunswik lens to investigate the cues related to ability and ability judgments. GMA and verbal ability were primarily related to verbal behavior, although they were both related to a few nonverbal and paraverbal behaviors. For example, verbal ability was positively related to speech rate. Overall, interviewers tended to ignore more valid cues and use more invalid cues than did the ML models. The GMA and verbal ability ML models had higher achievement and knowledge (in Brunswik lens parlance) compared to the intellect ratings, meaning they used cues more validly than interviewers. Additionally, the ML models were more consistent in their cue utilization compared to interviewers. However, as mentioned above, this did not prevent the models trained on intellect from being as valid in many aspects as the GMA and verbal ability predictions.

Theoretical Implications

The present study contributes to our understanding of the types of behaviors caused by ability, particularly in the interview context. Several findings support past theorizing regarding the relationship between ability and behavior. For example, personal pronouns, pronouns, and auxiliary verbs were negatively related to GMA and verbal ability. Additionally, analytical thinking, word count, and words longer than six letters were positively related to GMA and verbal ability. Pennebaker et al. (2014) argued that analytical thinking is related to ability and includes a negative weight for personal pronouns, impersonal pronouns, and auxiliary verbs, aligning well with the current findings. Additionally, Pennebaker and King (1999) and Robinson et al. (2013)

argued for the relationships that readability (e.g., FORCAST), word count, and long words have with ability and ability proxies. Observing these relationships in the present context speaks to a somewhat universal pattern of language related to ability across contexts. However, there were also findings specific to this study that do not reflect prior theorizing. For example, quantifiers were used more the higher one's GMA and verbal ability, while perceptual processes were talked about less. Something that LIWC cannot measure—the extent to which one speaks about relevant experiences—appeared to be reflected with positive correlations for both GMA and verbal ability with *n*-grams relevant to the interview (e.g., *project, work on*, Topic 15) and involved less repetition of the words and phrases in the questions themselves (e.g., *learn someth new,* Topic 25). Further, a present focus was negatively related to GMA and verbal ability, as was a future focus to verbal ability, perhaps because the interview consisted largely of past behavioral questions that should elicit past-focused responses. Aligning with these findings, Pennebaker and King (1999) found that present-focused speech was negatively related to ability. It seems that greater ability leads to more complex, abstract thinking that is reflected in speech, as well as more appropriate and informative responses—results which, while not surprising, speak to the advantages granted to interviewees of higher ability.

Construct validity is a persistent concern in employment interviews because method variance often contaminates the various constructs purportedly assessed (Hamdani et al., 2014). For example, interviewer ratings of intellect in the present study correlated r = .66 with conscientiousness ratings, yet even in selection settings where faking is likely present, conscientiousness only correlates ρ = .13 with ability (Schilling et al., 2021). One reason why construct discrimination in interviews may be poor is that ability is correlated with interview performance about as strongly as it is with job performance (Roth & Huffcutt, 2013). If, regardless of past experiences and suggested by the present study's findings, interviewees with greater ability provide better answers, then interview ratings of other constructs are inherently contaminated with ability. If the influence of ability on behavior can be thoroughly studied in the interview, one possibility for the future of AVIs is to partial out the variance specific to ability in order to isolate variance relevant to the focal KSAOs important for the job. Doing so could help with developing interviews that score multiple, distinct KSAOs, but it will require more theoretical development regarding the influence of ability on interview performance.

Overall, ability was primarily related to verbal behavior in the interview, not paraverbal or nonverbal behaviors. It may be that the actions of recalling events or sitting under the watchful eye of a camera elicits similar paraverbal and nonverbal behaviors regardless of one's ability. Indeed, it has long been suggested that there are many universal and habitual facial expressions, such as raising one's eyebrows when trying to recall some fact or occurrence or depressing the corners of one's mouth when anxious (Darwin, 1872; Duchenne, 1862). However, there were a few paraverbal and nonverbal behaviors that were related to ability. For example, jitter and shimmer, which are measures of low voice quality, were both negatively related to GMA and verbal ability. However, jitter and shimmer are sensitive to the tools and technologies used to capture the audio (Maryn et al., 2009). Considering this and the low strength of the relationships, this finding may be specific to the present study or even anomalous. On the other hand, finding that verbal ability related positively to speech rate aligns with prior research (Borkenau & Liebler, 1995; Reynolds & Gifford, 2001).

Regarding nonverbal behaviors, action unit 14, the chin raiser, was negatively related to both GMA and verbal ability. The chin raiser gives the appearance of a frown, a common expression during extended periods of concentration (Darwin, 1872) and during feelings of anxiety and uncertainty (Bitti et al., 2014; Ozel, n.d.). Alternatively, the chin raiser can also be activated during controlled smiles (Matsumoto & Willingham, 2009), which may be more evident for interviewees who are less able to present natural looking facial expressions. Each theoretical explanation for the relationship between the chin raiser and ability during interviews aligns with the idea that greater ability gives advantages in managing the simultaneous attentional demands inherent during interviewing, since greater ability can afford one greater control over one's self-presentation while responding to novel stimuli.

Regarding the Brunswik lens model, the present study raised some concerns about the theoretical underpinnings regarding cue utilization. Specifically, just because judgments are correlated with a behavior, does not mean that the judge actually used that behavior to make the judgment. For example, both GMA and verbal ability predictions from the combined model (i.e., LIWC Plus and *n*-grams) were correlated with inter utterance duration mean, a paraverbal behavior, even though the combined models did *not* include paraverbal behaviors as predictors. Clearly, the issues related to vicarious functioning—that when cues are highly intercorrelated, judges will have

difficulty validly choosing between them—also can affect whether the Brunswik lens model can provide clear insights regarding cue utilization.

Practical Implications

Automated interview research has extensive practical considerations since the COVID-19 pandemic accelerated the use of video interviewing, whether automatic or not. Ensuring that such methods are reliable, valid, and unbiased is necessary to justify their adoption and avoid their being characterized as AI snake oil (e.g., Narayanan, 2019).

Prior, similar research has considered internal convergence that meets or exceeds single rater reliabilities as adequate evidence to justify adopting ML systems to replace one or more human raters (Campion et al., 2016; Hickman, Bosch, et al., 2021). Attaining adequate convergence with the ground truth matters because it likely affects whether the correlations between the ground truth and other measures (e.g., job performance) pass through to the predictions. In the present study, single rater one-way random intraclass correlations, often referred to as ICC(1) and/or ICC(1, 1), were .27 for intellect ratings and .44 for extraversion and hireability ratings. Internal convergence for GMA models based on verbal behavior was slightly larger than the single rater reliability of intellect ratings, and convergence for verbal ability models was slightly lower than the single rater reliability of hireability ratings. Pairing this evidence with the fact that ML predictions exhibited attenuated external correlations compared to the observed test scores suggests that higher internal convergence is likely needed to justify adopting AVIs trained to model ability test scores. More reliable ability tests and/or larger training sample sizes may improve convergence in future studies. On the other hand, internal convergence for the intellect models far exceeded the single rater reliability of intellect ratings and were also larger than single rater reliability of hireability ratings. Additionally, the intellect models, at times, exhibited increased external convergent correlations compared to the observed scores, suggesting that convergence of .49 can be adequate for maintaining the validity of the ground truth measure. This evidence suggests that AVIs trained to model intellect ratings could be used to standardized and supplement pre-hire assessments.

The consideration of adequate convergence with the ground truth suggests that the overall validity of ML-based assessments rests on two key factors. First, the validity of the ground truth measure—a more reliable and valid ground truth should improve the psychometric properties of

the predicted values. Second, how accurately the ground truth measure is modeled by the ML-based assessment. Therefore, AVIs should both a) utilize reliable and valid measures as ground truth, and b) exhibit adequate convergence with ground truth measures.

These patterns of validity evidence suggest the importance of going beyond accuracy in predicting the ground truth (i.e., internal convergence) when evaluating ML-based assessments. For example, many studies of AVIs go no further than investigating how accurately the ground truth was modeled, either via error rates (e.g., root mean squared error) or correlational accuracy (e.g., Pearson's *r*; R^2; for exceptions see Hickman, Bosch, et al., 2021; Naim et al., 2018). Such investigations provide only one small piece of evidence regarding the validity of AVIs, as convergence with the ground truth represents just the first step in evaluating ML-based assessments. The broader nomological network of ML-based assessments should be investigated, including evaluating convergence with other, similar measures and whether AVIs adequately discriminate among multiple constructs. Additionally, before deploying AVIs for selection, their criterion evidence of validity should be established.

In the present study, models based on verbal behavior tended to be much more accurate than models based on paraverbal or nonverbal behavior (Table 6). These findings align with public sentiment and emerging practices in the field. Concerns have been raised to the Federal Trade Commission and Equal Employment Opportunity Commission regarding the legality of using facial recognition software to analyze nonverbal behavior (e.g., EPIC, 2018; Harris et al., 2018). Additionally, some vendors have emerged marketing chat-based automated interviews that use only text for analysis (Jayaratne & Jayatilleke, 2020), and some vendors are touting the use of only verbal behavior as predictors as more valid and defensible than including paraverbal and/or nonverbal behaviors (Caprino, 2021). As NLP methods continue to develop, language-based models will continue to grow in accuracy, and adding paraverbal and nonverbal behaviors appear to contribute little to the prediction of ability (although this may not be the case for all KSAOs).

More broadly, ability primarily affects what interviewees say in an interview, but interviewers tend to use several paraverbal and nonverbal behaviors that may contaminate their assessments of interviewee ability. Such effects may persist regardless of the KSAO being assessed (e.g., DeGroot & Motowidlo, 1999). These findings illustrate the importance of following structured interview protocols with behaviorally anchored rating scales to help reduce the potential for interviewer biases and subjectivity to contaminate interview ratings (Campion et al., 1997).

Limitations and Future Work

Several limitations in the present study inspire directions for future research. First, the present study drew on the Brunswik lens model to analyze aggregate judgments, as is commonly done in social psychology (e.g., Gifford, 1994). Therefore, the present application of the Brunswik lens model did not include idiographic investigations of individual raters. However, future research could benefit from exploring whether there is value in generating separate predictive models for each rater, as would traditionally be done during bootstrapping (Karelaia & Hogarth, 2008). Doing so, then aggregating scores from the separate models, may improve validity similar to how generating ML models for each item in a self-report scale and aggregating their scores can increase convergence (Hall & Matz, 2020). Additionally, this may help to mitigate problems caused by the inclusion of raters who are inconsistent in the application of their judgments, whether due to random error or bias. For example, lower quality raters may reduce both the reliability (i.e., intraclass correlations) and validity of ratings. By creating a model of their judgments, their judgment policies can be applied uniformly across participants, which may also enhance fairness.

Second, although automatic transcription enhances ecological validity, doing so likely attenuates construct validity of the ML models. At times, visual inspection of transcripts makes it clear that many errors are being introduced when using automatic transcription, although some transcripts appear very accurate. This is a thorny issue for vendors of AVIs as well because transcript quality is likely associated with the quality of computer hardware and internet connection used, which may be systematically associated with socioeconomic status. In other words, the use of automatic transcription may cause systematic error, or bias, associated with socioeconomic status. This could systematically disadvantage the disadvantaged, yet we know little about how transcription accuracy relates to individual differences or AVI validity in ecologically valid contexts. Early automated video-based assessment research found that computerized transcription harmed validity (Biel et al., 2013), yet more recent research found that it did not (Muralidhar et al., 2018). However, that research was based on video resumes scrubbed from YouTube, so future work should investigate these effects for one-way interviews.

Third, and related to the previous point, algorithmic bias continues to be a concern of ML-based assessments. Mean-level differences in and of themselves do not indicate bias, yet any exacerbation of group differences by the ML models likely does. In the combination models' predictions, Black-White differences were exacerbated slightly for all three operationalizations of

ability ($\overline{\Delta d}$ = -.05), yet since Hispanics and Indians made gains relative to Whites in the predicted values, the Black-Hispanic ($\overline{\Delta d}$ = -.26) and Black-Indian ($\overline{\Delta d}$ = -.25) differences grew about five times more than the Black-White differences. However, these differences are specific to the algorithm used—Tables A1 and A2 report the same information in Tables 11 and 12 but for the random forest combination models. Black-White differences in the random forest predicted values *decreased* compared to the observed values ($\overline{\Delta d}$ = .11). This suggests the importance of high inter-algorithm reliability for obtaining similar results across trained models. Differing levels of accuracy across ethnicity may have contributed to these changes in the magnitude of mean differences, although no convergent correlations were significantly different in the present study. In the future, model selection should be driven by more than just convergence with the ground truth, and bias is one important consideration. Using ML for assessment holds the potential to enhance fairness and reduce bias (Kleinberg et al., 2018), yet work in this area is still in its preliminary stages (e.g., Yan et al., 2020). More research is needed on investigating legal, robust, and replicable methods of debiasing ML-based assessments. For example, it was recently proposed that creating balanced, matched samples of majority and minority interviewees in the training sample may help to reduce bias (Tay et al., 2021). However, it is unclear whether such techniques can be effective for addressing bias across more than two demographic groups.

Fourth, the use of a non-applicant sample limits the study's ecological validity. Compared to applicants, non-applicants may be less likely to provide responses that reflect their maximum performance. If so, then using applicant samples will likely enhance the resulting ML models' validity since both ability test scores and interview performance will reflect high effort responses that reflect the interviewee's maximum performance. Some other elements of the study may have also limited the observed validity of the ML models, including the relatively short ability tests, mock interview, that the interview did not have multiple questions focused on eliciting ability, and the relatively small sample size (cf. Jayaratne & Jayatilleke, 2020). Similarly, using more situational instead of past behavioral questions may improve validity since some have argued that situational interviews are often merely verbally administered ability tests (Hunter & Hirsch, 1987), and some evidence supports the idea that situational interview scores are more strongly related to ability than past behavior interview scores (e.g., Day & Carroll, 2003; Kluemper et al., 2015). Using a sample of students also prevented the investigation of criterion evidence of validity—evidence that is necessary to justify automated interview ability assessments for selection.

CONCLUSION

The present study investigated the reliability and validity of algorithmic ability assessments in video interviews. Overall, predictions of GMA, verbal ability, and intellect had very similar patterns of relationships with proxies for ability, although intellect predictions had both higher reliability and worse construct discrimination compared to GMA and verbal ability predictions. Ability primarily affected what interviewees said in an interview, but interviewers used several paraverbal and nonverbal behaviors that may contaminate their assessments of interviewee ability. Such information and further work in this vein can help advance the reliability and validity of both algorithmic and traditional, face-to-face employment interviews.

REFERENCES

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219-245. https://doi.org/10.1037//0033-2909.121.2.219

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150-159. https://doi.org/10.1016/j.paid.2017.12.018

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology, 94*(6), 1394-1411. https://doi.org/10.1037/a0016532

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 1-4.

Biel, J.-I., Tsiminaki, V., Dines, J., & Gatica-Perez, D. (2013). Hi YouTube! Personality impressions and verbal content in social video. In *International Conference on Multimodal Interaction (ICMI'13)* (pp. 119-126). Sydney, Australia: ACM. https://doi.org/10.1145/2522848.2522877

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual levels of subnormals. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (Reprint, pp. 90-96). New York, NY: Appleton-Century-Crofts (Reprinted in 1961).

Bitti, P. E. R., Bonfiglioli, L., Melani, P., Caterina, R., & Garotti, P. (2014). Expression and communication of doubt/uncertainty through facial expression. *Ricerche di Pedagogia e Didattica. Journal of Theories and Research in Education, 9*(1), 159-177.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

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

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*(3), 546-553. https://doi.org/10.1037/0022-3514.65.3.546

Borkenau, P., & Liebler, A. (1995). Observable attributes as manifestations and cues of personality and intelligence. *Journal of Personality*, *63*(1), 1-25.

Bosch, N., & D'Mello, S. K. (in press). Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2019.2908837

Bourdage, J. S., Roulin, N., & Tarraf, R. (2018). "I (might be) just that good": Honest and deceptive impression management in employment interviews. *Personnel Psychology, 7*(4), 597-632. https://doi.org/10.1111/peps.12285

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50*(3), 255-272.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin*, *49*(6), 654-656.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco, CA: Jossey-Bass.

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*(7), 958-975. https://doi.org/10.1037/apl0000108

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

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

Caprino, K. (2021). *How AI can remove bias from the hiring proces and promote diversity and inclusion*. Available at https://www.forbes.com/sites/kathycaprino/2021/01/07/how-ai-can-remove-bias-from-the-hiring-process-and-promote-diversity-and-inclusion/

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Caylor, J. S. & Sticht, T. G. (March, 1973). Development of a simple readability index for job reading material. *Paper presented at the Annual Meeting of the American Educational Research Association*. Available at https://files.eric.ed.gov/fulltext/ED076707.pdf

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2017). The datafication of talent: How technology is advancing the science of human potential at work. *Current Opinion in Behavioral Sciences*, *18*, 13-16. https://doi.org/10.1016/j.cobeha.2017.04.007

Chamorro-Premuzic, T., & Furnham, A. (2005). *Personality and intellectual competence*. Mahwah, NJ: Lawrence Erlbaum Associates.

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, 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

Coleman, E. B. (1971). Developing a technology of written instruction: Some determiners of the complexity of prose. In E. Rothkopf & P. J. Rothkopf (Eds.), *Verbal learning research and the technology of written instruction* (pp. 155-204). New York, NY: Columbia University.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092-1122. https://doi.org/10.1037/a0021212

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper & Row.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14-27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Darwin, C. (1872). *The expression of emotions in animals and man*. London, England: John Murray.

Day, A. L., & Carroll, S. A. (2003). Situational and patterned behavior description interviews: A comparison of their validity, correlates, and perceived fairness. *Human Performance*, *16*(1), 25-47. https://doi.org/10.1207/S15327043HUP1601_2

DeGroot, T., & Motowidlo, S. J. (1999). Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology, 84*(6), 986-993.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61-84.

Duchenne, G. B. (1862). *Album de photographies pathologiques: complémentaire du livre intitulé De l'électrisation localisée*. J. B. Baillière et fils.

Dunn,W. S., Mount, M. K., Barrick, M. R., & Ones, D. S. (1995). Relative importance of personality and general mental ability in managers' judgments of applicant qualifications. *Journal of Applied Psychology, 80*, 500-509.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: a technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Emotient. (2015). Emotient SDK. Available at https://github.com/matteosimone/emotient-python

Eyben, F. (2014). *Real-time speech and music classification by large audio feature space extraction*. https://doi.org/10.1007/978-3-319-27299-3

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

Feiler, A. R., & Powell, D. M. (2016). Behavioral expression of job interview anxiety. *Journal of Business and Psychology*, *31*(1), 155-171. https://doi.org/10.1007/s10869-015-9403-z

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54.

Fernández-Martínez, F., Zablotskaya, K., & Minker, W. (2012). Text categorization methods for automatic estimation of verbal intelligence. *Expert Systems with Applications*, *39*(10), 9807-9820.

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*(2), 296-321. https://doi.org/10.1037/a0026556

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology*, *66*(2), 398-412. https://doi.org/10.1037/0022-3514.66.2.398

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological science*, *30*(7), 1087-1096.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*, 422-432.

Gonzalez-Mulé, E., Carter, K. M., & Mount, M. K. (2017). Are smarter people happier? Meta-analyses of the relationships between general mental ability and job and life satisfaction. *Journal of Vocational Behavior*, *99*, 146-164. https://doi.org/10.1016/j.jvb.2017.01.003

Gonzalez-Mulé, E., Mount, M. K., & Oh, I. S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology*, *99*(6), 1222-1243. https://doi.org/10.1037/a0037547

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504-528. https://doi.org/10.1016/S0092-6566(03)00046-1

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24 1 SPEC.*, 79-132. https://doi.org/10.1016/s0160-2896(97)90014-3

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science, 13*(1), 1-4. https://doi.org/10.1111/j.0963-7214.2004.01301001.x

Griffin, B. (2014). The ability to identify criteria: Its relationship with social understanding, preparation, and impression management in affecting predictor performance in a high-stakes selection context. *Human Performance, 27*(2), 147-164. https://doi.org/10.1080/08959285.2014.882927

Guiraud, P. (1954). *Les caracte`res statistiques duvocabulaire*. Presses universitaires de France

Hall, A. N., & Matz, S. C. (2020). Targeting Item-level Nuances Leads to Small but Robust Improvements in Personality Prediction from Digital Footprints. *European Journal of Personality, 34*(5), 873-884. https://doi.org/10.1002/per.2253

Harwell, D. (2019). A face-scanning algorithm increasingly decides whether you deserve the job. Retrieved from https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Hathaway, S., & McKinley, J. (1943). *Manual for administering and scoring the MMPI*. Minneapolis, MN: National Computer Systems.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*(3), 639-683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, in press, 1-82. https://doi.org/10.1037/apl0000695

Hickman, L., Saef, R., Ng, V., Tay, L., Woo, S. E., & Bosch, N. (2021). Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal*, in press.

Hornik, K., & Grün, B. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1-30.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1&2), 152-194. https://doi.org/10.1111/1468-2389.00171

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.

Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper, & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 321-357). Chichester, United Kingdom: Wiley.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594-612. https://doi.org/10.1037/0021-9010.91.3.594

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review*, *71*(1), 42-60.

Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, *30*(2), 387-398. https://doi.org/10.1007/s10869-014-9368-3

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102-138). New York, NY: Guilford Press. https://doi.org/10.2307/1213263

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621-652.

Judge, T. A., Ilies, R., & Dimotakis, N. (2010). Are health and happiness the product of wisdom? The relationship of general mental ability to educational and occupational attainment, health, and well-being. *Journal of Applied Psychology*, *95*(3), 454-468. https://doi.org/10.1037/a0019084

Judge, T. A., & Zapata, C. P. (2015). The person-situation debate revisited: Effect of situation strength and trait activation on the validity of the big five personality traits in predicting job performance. *Academy of Management Journal*, *58*(4), 1149-1179. https://doi.org/10.5465/amj.2010.0837

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404-426. https://doi.org/10.1037/0033-2909.134.3.404

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113-174. https://doi.org/10.1093/jla/laz001

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work. *Organizational Psychology Review, 1*(2), 128-146. https://doi.org/10.1177/2041386610387000

Kleisner, K., Chvátalová, V., & Flegr, J. (2014). Perceived intelligence is associated with measured intelligence in men but not women. *PLoS ONE, 9*(3). https://doi.org/10.1371/journal.pone.0081237

Kluemper, D. H., McLarty, B. D., Bishop, T. R., & Sen, A. (2015). Interviewee selection test and evaluator assessments of general mental ability, emotional intelligence and extraversion: relationships with structured behavioral and situational interview performance. *Journal of Business and Psychology*, *30*(3), 543-563. https://doi.org/10.1007/s10869-014-9381-6

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment*, *15*(3), 283-292. https://doi.org/10.1111/j.1468-2389.2007.00388.x

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805. https://doi.org/10.1073/pnas.1218772110

Küfner, A. C. P., Back, M. D., Nestler, S., & Egloff, B. (2010). Tell me a story and I will tell you who you are! Lens model analyses of personality and creative writing. *Journal of Research in Personality*, *44*(4), 427-435. https://doi.org/10.1016/j.jrp.2010.05.003

Kutik, B. (2015). HireVue: From video to predictive analytics. Retrieved from http://hrearchive.lrp.com/HRE/print.jhtml?id=534359174

Lang, J. W. B., & Kell, H. J. (2019). General mental ability and specific abilities: Their relative importance for extrinsic career success. *Journal of Applied Psychology*, *105*(9), 1047-1061. https://doi.org/10.1037/apl0000472

Lang, J. W. B., Kersting, M., Hülsheger, U. R., & Lang, J. (2010). General mental ability , narrower cognitive abilities , and job performance: The perspective of the nested-factors model of cognitive abilities. *Personnel Psychology*, *63*, 595-640.

Lee, A. J., Hibbs, C., Wright, M. J., Martin, N. G., Keller, M. C., & Zietsch, B. P. (2017). Assessing the accuracy of perceptions of intelligence based on heritable facial features. *Intelligence*, *64*, 1-8. https://doi.org/10.1016/j.intell.2017.06.002

Liem, C. S., Langer, M., Demetriou, A. M., Hiemstra, A. M. F., Wicaksana, A. S., Born, M., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, , X. Baro, Y. Güclütürk, U. Güclü, & M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197-253). New York, NY: Springer.

Lievens, F., Highhouse, S., & De Corte, W. (2005). The importance of traits and abilities in supervisors' hireability decisions as a function of method of assessment. *Journal of Occupational and Organizational Psychology*, *78*(3), 453-470. https://doi.org/10.1348/096317905X26093

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635-694.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology*, *86*(1), 96-111. https://doi.org/10.1037/0022-3514.86.1.96

Matsumoto, D., & Willingham, B. (2009). Spontaneous facial expressions of emotion of congenitally and noncongenitally blind individuals. *Journal of Personality and Social Psychology*, *96*(1), 1-10.

McGovern, T. V., & Tinsley, H. E. A. (1978). Interviewer evaluations of interviewee nonverbal behavior. *Journal of Vocational Behavior*, *13*(2), 163-171. https://doi.org/10.1016/0001-8791(78)90041-6

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437-455. https://doi.org/10.1037/a0028085

Melchers, K. G., Klehe, U.-C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance, 22*(4), 355-374. https://doi.org/10.1080/08959280903120295

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*(3), 683-729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Muralidhar, S., Nguyen, L. S., Frauendorfer, D., Odobez, J. M., Mast, M. S., & Gatica-Perez, D. (2016). Training on the job: Behavioral analysis of job interviews in hospitality. *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 84-91. https://doi.org/10.1145/2993148.2993191

Muralidhar, S., Nguyen, L., & Gatica-Perez, D. (2018). Words worth: Verbal content and hireability impressions in YouTube video resumes. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 322-327. https://doi.org/https://doi.org/10.18653/v1/P17

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin*, *33*(3), 325-339. https://doi.org/10.1177/0146167206294871

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality*, *71*(3), 465-493. https://doi.org/10.1111/1467-6494.7103008

Murzintcev, N., & Chaney, N. (2015). ldatuning: Tuning of the latent dirichlet allocation models parameters. *R package version 0.2-0, URL https://CRAN.R-project.org/package=ldatuning*.

Naim, I., Tanveer, I., Gildea, D., & Hoque, M. E. (2018). Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, *9*(2), 191-204. https://doi.org/10.1109/TAFFC.2016.2614299

Narayanan, A. (2019). How to recognize AI snake oil. https://static1.squarespace.com/static/52d805bde4b09cce38a94ff9/t/5dd5c0080e89843019d9a4ae/1574289418156/MIT-STS-AI-snakeoil.pdf

Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology*, *58*(2), 367-408. Retrieved from http://proquest.umi.com/pqdweb?did=851499911&Fmt=7&clientId=12010&RQT=309&VName=PQD

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hireability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, *16*(4), 1018-1031. https://doi.org/10.1109/TMM.2014.2307169

Nguyen, L. S., & Gatica-Perez, D. (2016). Hireability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia*, *18*(7), 1422-1437. https://doi.org/10.1109/TMM.2016.2557058

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Ones, D. S., Kaiser, R. B., Chamorro-Premuzic, T., & Svensson, C. (2017). Has industrial-organizational psychology lost its way? *The Industrial-Organizational Psychologist, 54*(4). Retrieved from http: //www.siop.org/tip/april17/lostio.aspx

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1), 505-533. https://doi.org/10.1146/annurev-orgpsych-032117-104553

Ozel, M. (n.d.). *Lower face cheat sheet*. Available at https://melindaozel.com/

Ozer, D. J., & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology*, *45*, 357-388. https://doi.org/10.1016/B978-0-12-375000-6.00272-X

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

Paulhus, D. L., & Morgan, K. L. (1997). Perceptions of intelligence in leaderless groups: The dynamic effects of shyness and acquaintance. *Journal of Personality and Social Psychology*, *72*(3), 581-591. https://doi.org/10.1037/0022-3514.72.3.581

Peeters, H., & Lievens, F. (2006). Verbal and nonverbal impression management tactics in behavior description and situational interviews. *International Journal of Selection and Assessment*, *14*(3), 206-222. https://doi.org/10.1111/j.1468-2389.2006.00348.x

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Piedmont, R. L., McCrae, R. R., Rieman, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*(3), 582-593. https://doi.org/10.1037/0022-3514.78.3.582

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., & Escalera, S. (2016). Chalearn LAP 2016: First round challenge on first impressions - Dataset and results. *Lecture Notes in Computer Science*, *9915 LNCS*(October), 400-418. https://doi.org/10.1007/978-3-319-49409-8_32

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2019). Mitigating bias in algorithmic employment screening: Evaluating claims and practices. *ArXiv*, 1-24. https://doi.org/10.2139/ssrn.3408010

Reynolds, D. J., & Gifford, R. (2001). The sounds and sights of intelligence: A lens model channel analysis. *Personality and Social Psychology Bulletin*, *27*(2), 187-200. https://doi.org/10.1177/0146167201272005

Roth, P. L., & Huffcutt, A. I. (2013). A meta-analysis of interviews and cognitive ability: Back to the Future? *Journal of Personnel Psychology*, *12*(4), 157-169. https://doi.org/10.1027/1866-5888/a000091

Rotolo, C. T., Church, A. H., Adler, S., Smither, J. W., Colquitt, A. L., Shull, A. C., Paul, K. B., & Foster, G. (2018). Putting an end to bad talent management: A call to action for the field of industrial and organizational psychology. *Industrial and Organizational Psychology*, *11*(2), 176-219. https://doi.org/10.1017/iop.2018.6

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207-1225. https://doi.org/10.1037/apl0000405

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schilling, M., Becker, N., Grabenhorst, M. M., & König, C. J. (2021). The relationship between cognitive ability and personality scores in selection situations: A meta-analysis. *International Journal of Selection and Assessment*, *29*(1), 1-18.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262-274. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.172.1733

Schmidt, F. L., Shaffer, J. A., & Oh, I. S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, *61*(4), 827-868. https://doi.org/10.1111/j.1744-6570.2008.00132.x

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, *10*(1), 1-22. https://doi.org/10.1177/014662168601000101

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, *25*(1), 12-27. https://doi.org/10.1016/j.hrmr.2014.09.004

Sergienko, R., & Schmitt, A. (2015). Verbal intelligence identification based on text classification. In *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2524-2528.

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* (5th ed.). Washington, DC: American Psychological Association. https://doi.org/10.1017/iop.2018.195

Sparck Jones, K. (1972). Statistical interpretation of term specificity and its implication in retrieval. *Journal of Documentation, 28*(1), 11-21.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-293.

Speer, A. B., Christiansen, N. D., Melchers, K. G., König, C. J., & Kleinmann, M. (2014). Establishing the cross-situational convergence of the ability to identify criteria: Consistency and prediction across similar and dissimilar assessment center exercises. *Human Performance*, *27*(1), 44-60. https://doi.org/10.1080/08959285.2013.854364

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality, 34*(5), 826-844.

Tay, L., Woo, S. E., Hickman, L., Booth, B., & D'Mello, S. (2021). *A conceptual framework for investigating and mitigating machine learning bias for psychological assessment*. Manuscript submitted for publication.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.

Tiam-Lee, T. J., & Sumi, K. (2017, September). Analyzing facial expressions and hand gestures in filipino students' programming sessions. In *2017 International Conference on Culture and Computing (Culture and Computing)* (pp. 75-81). IEEE.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, *71*(6), 528-530.

Van Iddekinge, C. H., McFarland, L. A., & Raymark, P. H. (2007). Antecedents of impression management use and effectiveness in a structured interview. *Journal of Management*, *33*(5), 752-773. https://doi.org/10.1177/0149206307305563

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology*, *90*(3), 536-552. https://doi.org/10.1037/0021-9010.90.3.536

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281-300. https://doi.org/10.1037/a0017908

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of g-theory methods for modeling multitrait-multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*(1), 134-161. https://doi.org/10.1177/1094428111408616

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platan, C. Ma, Y. Jernite, J. Plu, C. Xu, T. C. Scao . . . Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38-45). Stroudsburg, PA: Association for Computational Linguistics

Yankov, G. P., Wexler, B., Haidar, S., Kumar, S., Zheng, J., & Li, A. (2020). *Algorithmic justice*. SIOP White Paper Series.

Zablotskaya, K. (2015). *Automatic estimation of users' verbal intelligence* (Doctoral dissertation, Universität Ulm). Ulm, Germany.

Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, *28*(2), 238-249. https://doi.org/10.1177/0146167202282009

Zeigler-Hill, V., Besser, Y., & Besser, A. (2019). A negative halo effect for stuttering? The consequences of stuttering for romantic desirability are mediated by perceptions of personality traits, self-esteem, and intelligence. *Self and Identity*, *19*(5), 613-628. https://doi.org/10.1080/15298868.2019.1645729

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.1467-9868.2005.00527.x

APPENDIX A

Table 15. Ability Scores Analyzed by Race and Gender for Random Forest Combination Models

Observed Scores	East Asian Mean (*SD*)	Black Mean (*SD*)	Hispanic Mean (*SD*)	Indian Mean (*SD*)	White Mean (*SD*)	Women Mean (*SD*)	Men Mean (*SD*)
GMA	8.47 (3.26)	6.23 (2.95)	7.39 (3.25)	8.02 (3.28)	8.54 (3.06)	8.10 (3.17)	8.33 (3.34)
Verbal Ability	18.42 (7.56)	14.89 (6.93)	18.35 (8.37)	18.02 (8.15)	20.34 (7.26)	19.20 (7.54)	18.75 (7.65)
Intellect	5.49 (.61)	5.19 (.80)	5.33 (.59)	5.62 (.64)	5.50 (.67)	5.47 (.66)	5.46 (.68)
	East Asian-White Cohen's *d*	Black-White Cohen's *d*	Hispanic-White Cohen's *d*	Indian-White Cohen's *d*		Women-Men Cohen's *d*	
GMA	-.02	-.77	-.36	-.16		-.07	
Verbal Ability	-.26	-.77	-.25	-.30		.06	
Intellect	-.02	-.42	-.27	.18		.01	
Predicted Values	East Asian Mean (*SD*)	Black Mean (*SD*)	Hispanic Mean (*SD*)	Indian Mean (*SD*)	White Mean (*SD*)	Women Mean (*SD*)	Men Mean (*SD*)
GMA	8.14 (.42)	7.93 (.48)	8.13 (.47)	8.21 (.40)	8.21 (.49)	8.20 (.47)	8.09 (.47)
Verbal Ability	18.87 (1.13)	18.35 (1.30)	18.90 (1.24)	19.13 (1.08)	19.16 (1.24)	19.10 (1.21)	18.78 (1.20)
Intellect	5.44 (.15)	5.40 (.17)	5.44 (.17)	5.46 (.15)	5.4 (.17)	5.46 (.16)	5.43 (.17)

Table 15 continues

	East Asian-White Cohen's *d*	Black-White Cohen's *d*	Hispanic-White Cohen's *d*	Indian-White Cohen's *d*	Women-Men Cohen's *d*
GMA	-.15	-.58	-.17	.00	.23
Verbal Ability	-.24	-.64	-.21	-.03	.27
Intellect	-.19	-.41	-.18	-.06	.18

Note. East Asian *N* = 171; Black *N* = 56; Hispanic *N* = 54; Indian *N* = 44; White *N* = 364; Women *N* = 465; Men *N* = 262.

Table 16.Correlational Accuracy of Ability Predictions by Race and Gender for Random Forest Combination

	GMA	Verbal Ability	Intellect
Overall	.29	.40	.54
East Asian	.24	.36	.52
Black	.35	.47	.58
Hispanic	.14	.44	.52
Indian	.33	.30	.41
White	.29	.35	.56
Women	.32	.41	.56
Men	.27	.33	.52

Note. No correlations are significantly different at $p < .05$ using Fisher's *r*-to-*z* transformation, but power is very limited for all ethnic comparisons.

APPENDIX B

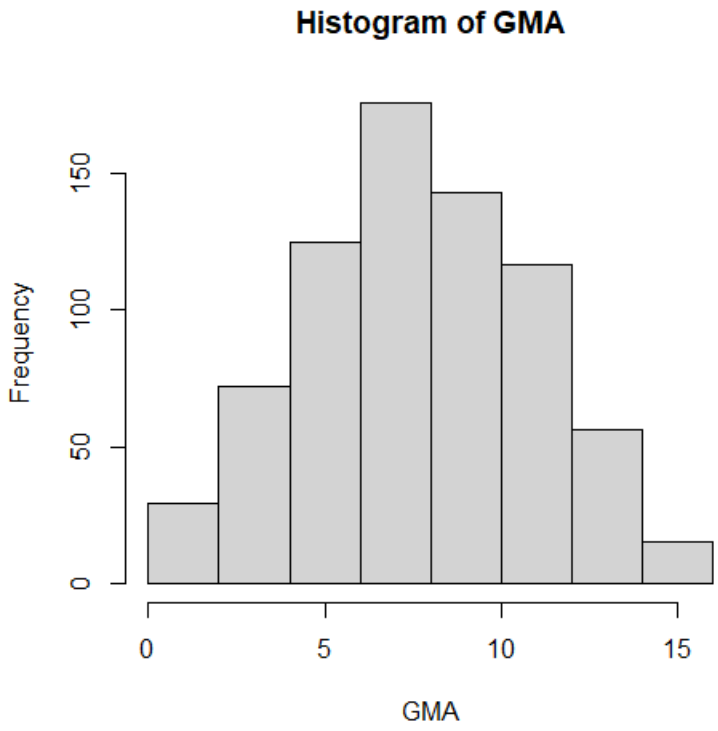


Figure 9. Histogram of GMA test scores in Sample 1.

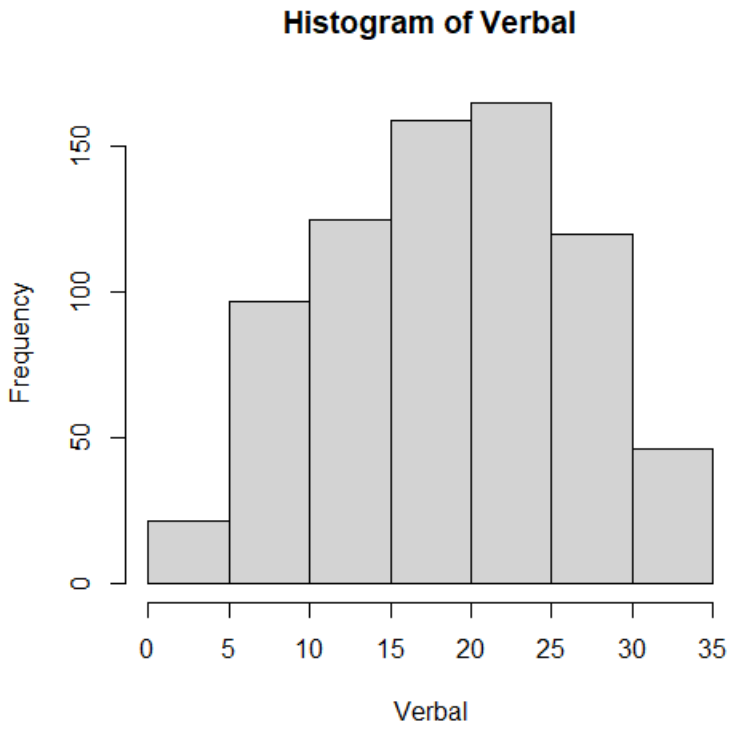


Figure 10. Histogram of verbal ability test scores in Sample 1.