HireVue

WHITE PAPER

HIREVUE'S ASSESSMENT SCIENCE

AUTHORS

Dr. Kiki Leutner: Director, Assessment Innovation

Dr. Josh Liff: Director, Assessment Psychometrics and Applied Research

Dr. Lindsey Zuloaga : Chief Data Scientist

Dr. Nathan Mondragon: Chief I/O Psychologist

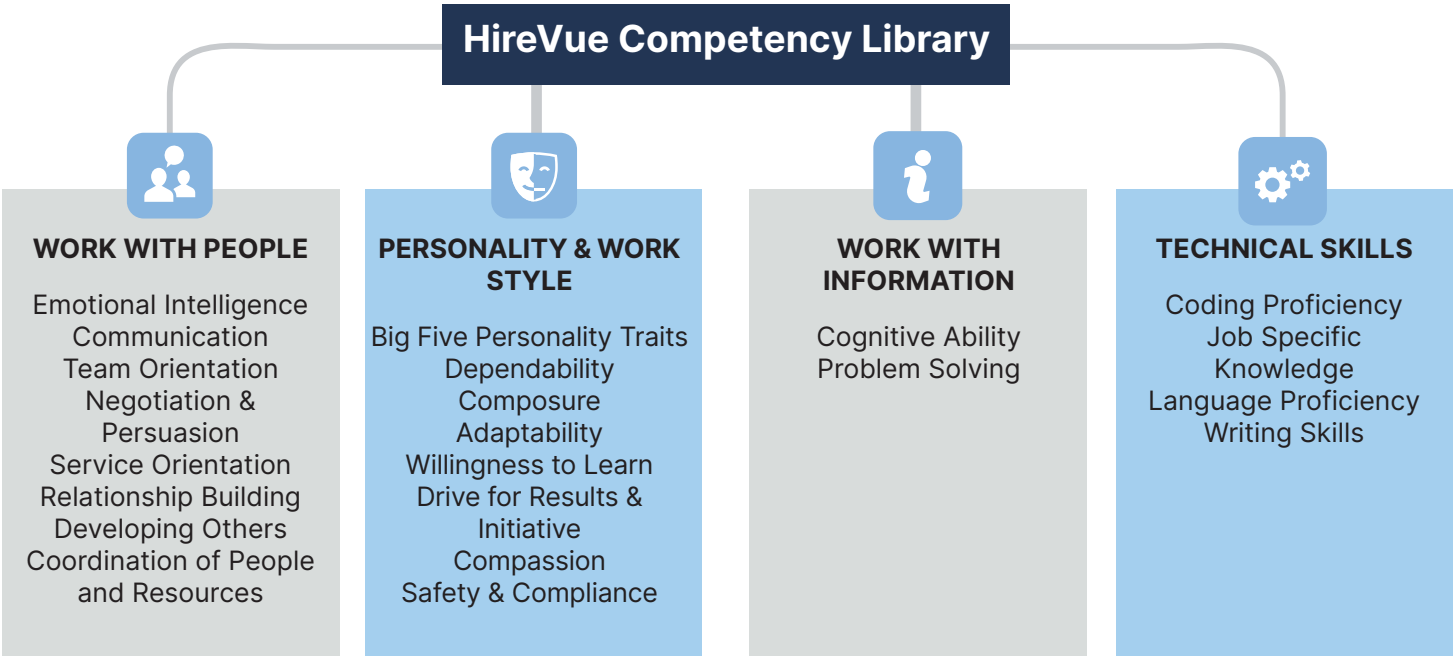
JUNE / 2021

TABLE OF CONTENTS

SUMMARY

HireVue assessments are **psychometric assessments** that measure traits and competencies associated with performance at work. Psychometric assessments and structured interviews are the most scientifically valid methods for making selection decisions (Schmidt & Hunter, 1998; Levashina, Hartwell, Morgeson, & Campion, 2014). They substantially reduce sources of human bias and result in a more diverse and skilled workforce (Ployhart, 2000; Ployhart, & Holtz, 2008). Decades of research show that there are a number of traits and competencies that are predictive of performance at work, including cognitive ability and personality (Barrick & Mount, 1991; Kuncel, Ones, Sackett, 2010; Gonzalez-Mule, Mount, & Oh, 2014). HireVue's scientific model summarizes these competencies into four key areas: working with people, personality and work style, working with information, and technical skills.

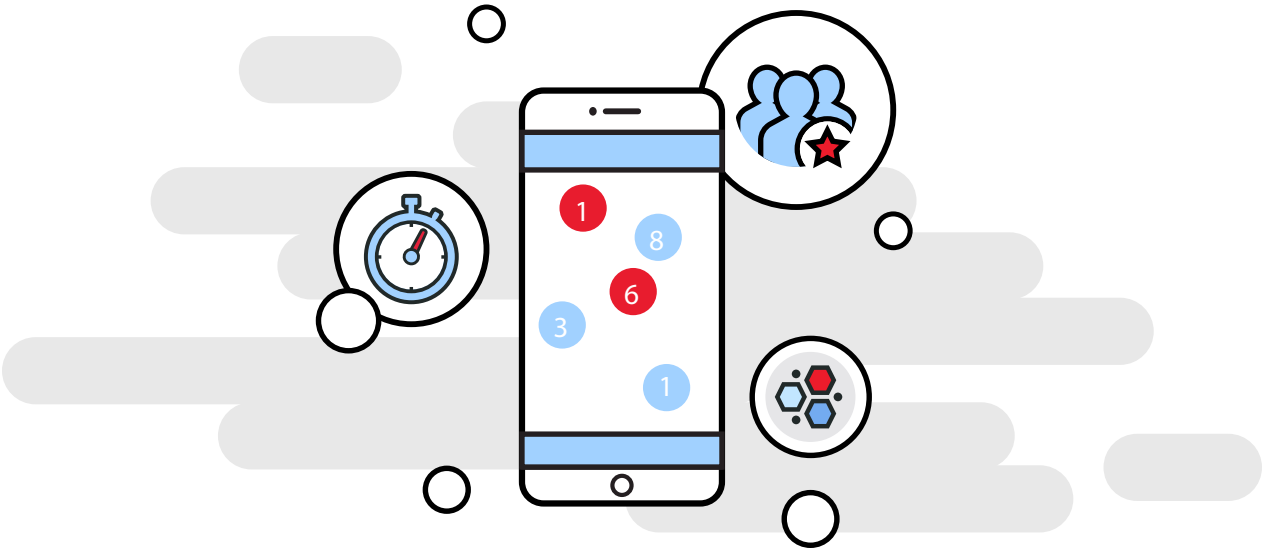
FIGURE 1: HIREVUE JOB FIT FRAMEWORK



There are well-established processes for developing and validating psychometric assessments for use by Talent Management and People Analytics professionals. All psychometric assessments, including many legacy tests such as self-report questionnaires, use algorithms to derive personality, cognitive ability, or competency scores for individual test-takers. As a result, and contrary to other domains where Artificial Intelligence is driving innovation, psychometric assessments are well regulated. They must meet strict standards of fairness and quality, which in the United States are posed by the Uniform Guidelines on Employee Selection Procedures (1978) as adopted by the US Equal Employment Opportunity Commission, and professional testing standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society for Industrial and Organizational Psychology, 2018). Our assessments are developed and monitored by an interdisciplinary team of **Data Scientists and Industrial and Organizational (I/O) Psychologists** who are guided by professional testing standards and regulatory bodies, as well as the application of principles and measures of algorithmic fairness throughout the assessment development and deployment process.

HireVue assessments introduce innovation to pre-hire assessments through novel formats: standardized video interviews and game-based assessments. These formats result in different data underlying the assessment scoring algorithm compared with traditional, questionnaire-based assessments. Apart from this, HireVue assessments are developed, validated, and used in the same way as traditional psychometric tests. All HireVue assessments measure traits and competencies found to be relevant in the workplace across a variety of roles and industries. They are mitigated for adverse impact to improve fairness, and meet strict standards of reliability and validity.

- **Interview assessments have convergent validities** of r = .51 to r = .62 with trained evaluator ratings of the respective competency.
- **Interview assessments have predictive validities** of r = .25 to r = .49 with various job- related outcomes**.**
- **Cognitive game-based assessments** correlate between r = .51 and r = .67 with a legacy cognitive ability test (ICAR; Condon & Revelle, 2014). Emotional intelligence game-based assessments between r = .36 and r = .45 with traditional tests measuring related constructs (GERT-S; Schlegel & Scherer, 2016; STEM; MacCann & Roberts, 2008). And Personality game based assessments between r = .50 and .73 with legacy personality assessments (IPIP, Johnson, 20140..



What HireVue assessments measure

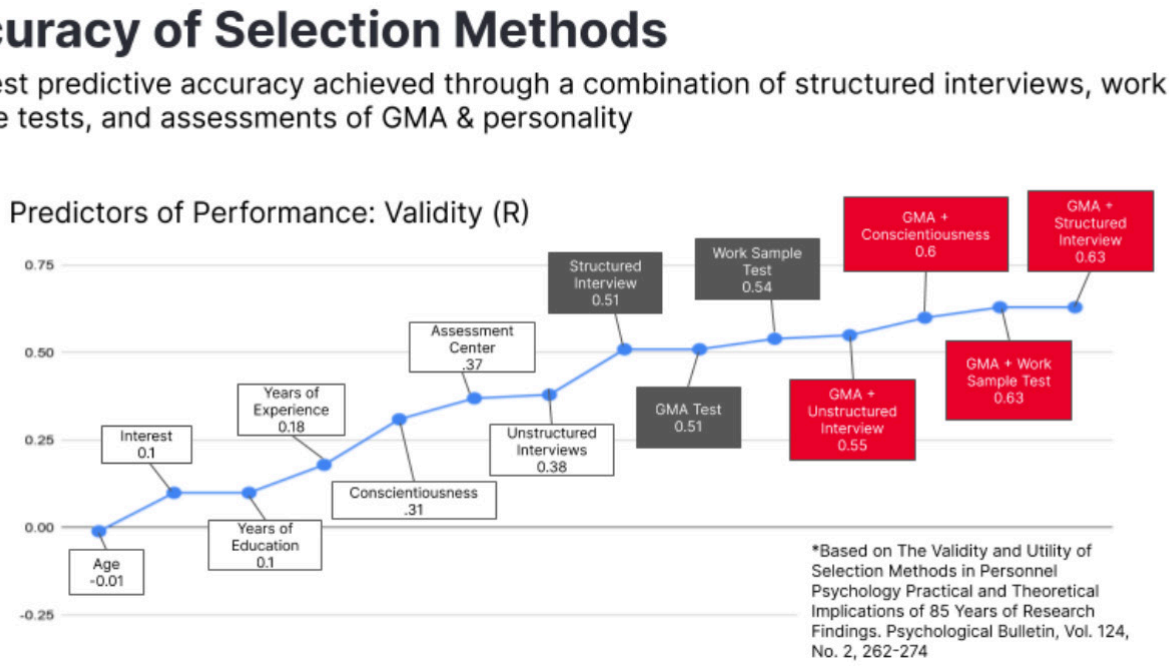
Traditional psychometric assessments have strong validity in predicting job performance (see Figure 2 and Table 1). Their use for talent management and recruiting as an estimate of talent and career potential dates back over 100 years (Kanfer, Ackerman, Murtha, & Goff, 1995; Ryan, & Ployhart, 2014). Indeed, few constructs in the social sciences have been as widely applied in practice as General Mental Abilities (GMA), job-relevant competencies measured via structured employment interviews, and personality traits for

the prediction of job performance and other career related outcomes (Schmitt, 2014; Schmidt, & Hunter, 1998). This research has consistently demonstrated that structured employment interviews are valid for predicting job performance criteria (Campion, Palmer, and Campion, 1997; McDaniel, Whetzel, Schmidt, and Maurer, 1994). Meta analysis of studies spanning 100 years demonstrates that structured interviews have some of the highest validity in predicting job-relevant outcomes compared with other common selection methods (Schmidt & Hunter, 1998; see Figure 2). This effect may be further enhanced by applying machine learning-based profiling to the already standardized video interview process. Human raters make inferences that are not job-related, bring idiosyncratic biases to the evaluation process, lack the ability to recall job-related details of the interview, and lack accuracy in making criterion-related judgments and decisions that predict success on the job

(Campion et al., 1997). By introducing automated scoring of candidate's text responses with machine learning, the fairness and standardization of human raters can be improved (Campion, Campion, Campion, & Reider; 2016). This results in greater efficiency and significant time savings, allowing employers to collect unstructured data and analyzing it in a structured and standardized manner.

FIGURE 2: PREDICTIVE VALIDITY OF DIFFERENT SELECTION METHODS FOR JOB PERFORMANCE

GMA= General Mental Ability



Both ability and soft skills are important for job performance

Candidates are typically evaluated by employers based on their ability to perform their job as well as their motivation (ambition, work ethic and drive) and interpersonal skills. While GMA is an important predictor of job success, a broader set of competencies and personality characteristics are needed to get a comprehensive assessment of employability and job fit (Hogan, Chamorro-Premuzic, & Kaiser, 2013). Indeed, based on one of the largest systematic meta-analysis of characteristics predicting job performance to date, GMA is the most consistent and strongest predictor of career success. Further, the combination of GMA and structured employment interviews provides incremental predictive validity over GMA alone (mean corrected validity of r = .63). These findings have been replicated in hundreds of studies over the past twenty years,including a meta-analytic review of 20,000 studies with over 5 million participants (Kuncel, Ones, & Sackett, 2010). In addition, personality traits also predict job performance beyond GMA (Sackett, Gruys, & Ellingson, 1998), and predict distinct aspects of job performance, demonstrating incremental validity over each other (Chamorro-Premuzic, & Arteche, 2008; Judge, Higgins, Thoresen, & Barrick, 1999; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Together, personality and GMA not only predict job performance (Dries, 2013) but also expertise (Ullén, Hambrick, & Mosing, 2016), leadership effectiveness (Cavazotte, Moreno, & Hickmann, 2012), ratings of employability, or job fit (Dunn, Mount, Barrick, & Ones, 1995) as well as competency proficiency level once in the role (Bartram, 2005). Collectively, GMA, structured employment interviews, and personality provide a more accurate prediction of job performance than any of the individual measures alone.

TABLE 1: PERSONALITY

Cognitive Ability and career success across job roles and industries. Specific traits have stronger or weaker relationships with job performance in different roles and industries. Based on Judge, Higgins, Thoresen & Barrick, 1999; Higgins, Peterson, Pihl & Lee, 2007.

	Supervisor rated performance (Judge et al., 1999)	**Extrinsic career success** (Higgins et al., 2007)
Cognitive ability	.42 – .57**	.53**
Extraversion	-.08 – .14	.19*
Agreeableness	.07 – .09	-.11
Conscientiousness	.08 – .23*	.50**
Emotional stability	.01 – .06	-.34**
Openness to experience	.03 – .13	.14

Machine learning as a new tool to assess people

Measuring the behavioral tendencies, preferences, and traits of individuals in a selection context has traditionally been done through self-report questionnaires. A wealth of studies show that we can detect personal characteristics in the behaviors of individuals. However, early research on the link between personality and behavioral observation required time-consuming manual extraction of behaviors from text, audio and video recordings (Mischel, 1996).

Digitalization and advances in machine learning have made these behaviors accessible to researchers in a structured way, leading to a wealth of publications showing the link between different behavioral data sources and personality. For example, language use on Twitter and personal blogs is indicative of one's traits (Schwartz et al., 2013) and Facebook likes are predictive of personality (Kosinski, Stillwell, & Graepel, 2013).

Generally, such approaches demonstrate good convergent validity with self-report measures of personality, and even outperformed human judges when comparing the convergence of inter- judge agreement for machine (r = .62) versus human judgments (r = .38) (Bleidorn & Hopwood, 2018). For example, machine learning algorithm predictions of Big Five traits - Openness, Conscientiousness, Agreeableness, Extraversion, and Neuroticism - are more accurate (r = .56 versus r = .49) than friends' judgments of Big Five traits (Youyou, Kosinski & Stillwell, 2015).

Behavior reveals tendencies and traits

A rich body of research demonstrates the link between language use and personality in particular (Kern et al., 2013; Pennebaker & King, 1999). The words we use are related to personal concerns and express the things we value (Chung & Pennebaker, 2014). The lexical approach, the idea that underlying psychological characteristics are embedded in the structure of language, has a long tradition in psychometrics and has been demonstrated in numerous studies (Gosling, Gaddis, & Vazire, 2007; Schwartz, Eichstaedt, Kern, et al., 2013; Vazire & Gosling, 2004).

The ability to mine and analyze large amounts of free text data has led to a surge in publications on semantics and personal characteristics (Lambiotte & Kosinski, 2014; Schwartz, Eichstaedt, Kern, et al., 2013). Language use-based measures show moderate correlations with the Big Five traits (r = .37 for Conscientiousness; Park et al., 2015). Language use-based measures may have advantages over self-report measures in that they require less effort from the test taker, and are less prone to response bias (Bardi, Calogero, & Mullen, 2008).

Nonverbal cues are also linked with personality traits, and personality judgments made by others. Speech signals such as rate, energy, pitch, and silent intervals successfully distinguish between high and low extraversion in 86% of cases (Kwon, Yeon Choeh, & Lee, 2013). In addition, audio and non-verbal behavior exhibited during interviews shows promising predictive validity in the organizational context. Automatically extracted behavioral cues from videos of interviewees and interviewers explain 36% of variance in hiring decisions, and are more predictive of hiring decisions compared to psychometric questionnaires (Nguyen, Frauendorfer, Mast, & Gatica-Perez, 2014).

Collectively, psychological studies using machine learning to predict personality characteristics to date show good convergent validities of digital footprints and behavioral observations with traditional measures.

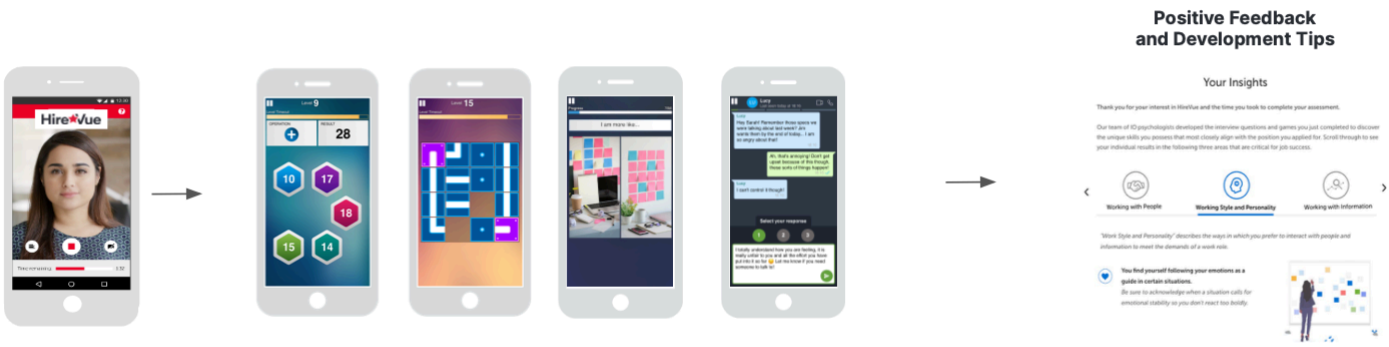
Interview and game based assesments

HireVue assessments are designed to collect data that is relevant to the employment context. Data captured in the assessment process is theoretically and empirically linked to job performance and to traits and competencies that are related to job performance. This data is collected in an engaging, fast, and transparent way through interviews and game-based assessments. Interview and game-based assessments may be combined into one assessment experience, as seen in an example of a typical HireVue assessment (Figure 3). Given their shorter length in comparison with traditional assessments, this allows for comprehensive testing of interpersonal skills, personality, and ability within one assessment session.

FIGURE 3:

A typical HireVue Assessment experience including interview and game-based components. Assessments can be taken at any time, in one go or with breaks in between.

A typical HireVue Assessment experience including interview and game-based components. Assessments can be taken any time, in one go or with breaks in between.



Ondemand interviews as psychometric assessments

Data collected during OnDemand interviews are used for HireVue's Interview Assessment scoring algorithms. OnDemand interviews have a set number of written or pre-recorded interview questions to which participants record their answers video or audio only. Typically, each trait or competency is assessed with one interview question. This ensures that candidates provide enough information to make a valid and reliable assessment. Questions are designed such that answers provide insights into a specific competency. For example, a question designed to measure Dependability is: *"Tell me about a time you had a challenge keeping a commitment to others at work or at school due to other priorities. Please describe the situation, your actions, and the outcome."*

DATA USED TO AUTOMATICALLY SCORE INTERVIEW ANSWERS AND DETERMINE JOB FIT

To build scoring algorithms for traits and competencies based on responses to interview questions, responses to interview assessment questions are transcribed from speech to text. A comprehensive set of features (i.e. variables) relating to response content are extracted to form the basis of a typical assessment.

There is an extensive body of literature showing the relationship between language use and personal characteristics, including personality traits and values (Pennebaker & King, 1999; Schwartz, Eichstaedt, Kern, et al., 2013, Yarkoni, T. 2010). All speech in an interview is transcribed using a transcription tool with state-of- the-art accuracy. These types of speech-to-text transcription algorithms are trained on diverse datasets to capture differences in local dialects of the respective language.

In addition to training speech-to-text transcription algorithms on diverse populations, once an interview-based scoring algorithm is deployed, HireVue has developed methods for detecting when poor quality audio is present, and then flags such cases for manual review. That is, factors that potentially impact the accurate scoring of an interview are monitored, and if detected, prevent the scoring of an interview. Factors responsible for causing low confidence in transcription accuracy may include: low volume audio, background noise, not enough words spoken, unintelligible voice and/or speaking in a different language than expected. When transcription confidence is extremely low, HireVue will not score the interview and generate an "Insufficient Data Error" alert in place of the final score to flag the interview for manual review.

Beyond raw words, variables relating to word choice, meaning, and sentiment are extracted using the General Inquirer taxonomy, generating over 400 variables (Loper & Bird, 2002; Stone, Dunphy & Smith, 1966). Variables label words as belonging to different categories, such as over or understatement, motivation words (goals, achievement, aspirations, etc.), or cognitive orientation (thinking, knowing, solving, etc.).

Another class of word features is created using part-of-speech (proper nouns, pronouns, adverbs, etc.) tagging with the Natural Language Toolkit (NLTK; Bird, Loper & Klein, 2009). Finally, in some newer models, features are engineered using the latest advancements in Natural Language Processing with a pre-trained algorithm called Robustly Optimized Bidirectional Encoder Representations from Transformers Pre-Training Approach (RoBERTa) that is fine-tuned on interview data. These features get at more nuance and context in speech and are robust to differences in word choice, focusing more on meaning and intention (Liu et al., 2019; Devlin, Chang, Lee & Toutanova, 2018).

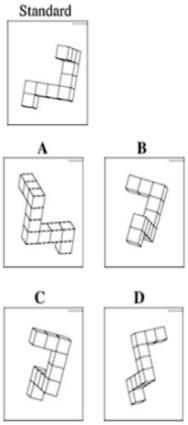
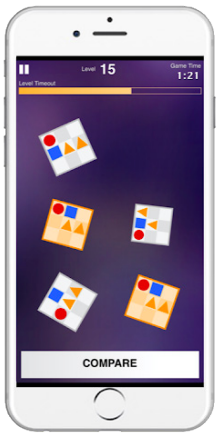
Spectral audio characteristics are extracted using a HireVue proprietary feature learning algorithm created specifically to extract audio features such as tonal variation and the length of gaps between utterances. These speech qualities have been shown to be indicative of personality and perceived emotional state (Mohammadi, G., Origlia, A., Filippone, M. & Vinciarelli, A., 2012, Sahu, G., 2019).

Games modeled after legacy psychometric tests

Games are designed to reflect legacy psychometric tests. For example, our game-based cognitive ability tests include games that are modeled after the quiz type questions typically included in legacy tests. A legacy test might include questions with patterns of shapes that need to be mentally rotated in order to identify a matching shape. With our game-based design, the same task is presented in a more engaging and fast-paced environment. The difficulty level adjusts based on the skill level of the candidate, with level complexity rising through the addition of shapes, colors, or distraction effects.

FIGURE 4:

Legacy versus game-based cognitive ability test.

LEGACY ASSESSMENT SAMPLE QUESTION:	***LEGACY ASSESSMENT***	***GAME-BASED ASSESSMENT***
+ WHICH FIGURE MATCHES THIS SHAPE?		

HireVue games are designed to assess specific traits or competencies. Data collected during the games are indicative of the players' characteristics in relation to the assessed trait. For example, cognitive ability games are designed such that each level increases in difficulty, and players progress to a higher level each time they complete a level successfully, or down if they fail. Therefore, data on the highest level completed, ratio of levels lost and won, and the total number of levels played is collected.

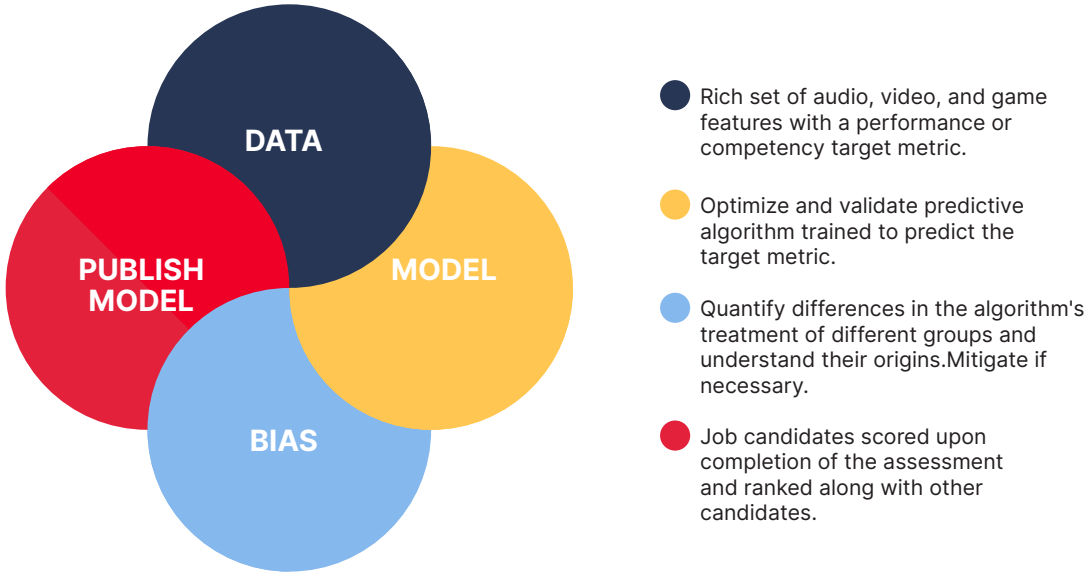
The advantages of using game-based assessments are an improved user experience, shorter testing times, and as a result, improved data quality (DeRight & Jorgensen, 2015; Miranda & Palmer, 2014). Gamified assessments decrease anxiety (Alter, Aronson, Darley, Rodriguez, & Ruble, 2010; McPherson & Burns, 2005) and increase engagement and motivation in test-takers (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012) through real-time feedback, advancement through levels, and clear goals (Burgers, Eden, van Engelenburg, & Buningh, 2015; Wood, Griffiths, Chappell, & Davies, 2004). Indeed, HireVue Game-Based assessments achieve 95% completion rates amongst applicants who start the assessment, and the HireVue Net Promoter Score is 70

Our validation process

In addition to innovation in the assessment format, interview and game assessments involve new scoring methods based on machine learning. These allow us to get information about the job-related competencies of the participants based on the large amounts of data that are collected during interview and game-based assessments.

FIGURE 5:

Assessment scoring algorithm validation process.



The scoring algorithms for each of the interview and game-based assessment components are developed in several steps:

1. A prediction objective is specified in which assessment data is used to predict the respective outcome score (e.g. scores on a traditional cognitive ability assessment for our cognition game based assessment, or ratings of a given competency for our interview assessments).

2. A range of suitable prediction models (for example Logistic Regression, Linear Regression, and Ridge Regression) are tested to determine the best performing model. Cross validation and regularization are used to control overfitting. These are best practices and are commonplace with machine learning applications. This step is to determine the optimal model and settings (hyperparameter values1) to use for training the final algorithm.

3. The top performing model is then tested for adverse impact and other metrics related to algorithmic fairness (see Section on Fairness and Adverse Impact for additional details). Where bias is identified, the model is mitigated such that top features contributing to the bias are either removed or de-weighted. After several iterations of this process, the model that shows the lowest adverse impact while maintaining predictive validity is selected as the scoring model for the assessment.

4. Once an assessment is live and in use by a customer, all incoming candidates are evaluated in the same way, regardless of their demographic class. The scoring models are monitored and updated to ensure that there is no adverse impact or scoring anomalies present at potential cut scores.

Convergent validity

To confirm that an assessment measures the job-related construct it is intended to measure, convergent validity is tested. The convergent validity evaluates the assessment's overlap with alternative measures of the trait or competency the assessment is measuring. Correlations are affected not only by the true relationship between the constructs measured, but also by variance resulting from the assessment method: where assessment methods are more similar, there is more shared variance, and the correlations will be stronger (Ventura & Shute, 2013; Wang, Shute, & Moore, 2015). Thus, the less similar the assessment formats, the lower the expected correlation (e.g. the Multi-trait, multi-method matrix, Campbell & Fiske, 1959). Therefore, correlations observed between new and traditional assessment scores are typically lower than correlations observed between two traditional questionnaires of the same trait.

Typical values we observe indicate the interview and game-based assessments overlap well with the traits they are intended to measure. For Interview Assessments measuring competencies, correlation coefficients range from r = . 51 to r = .62 depending on the trait assessed ($p < .001$, see Table 2). For game-based assessments of cognitive ability, r values vary between r = . 51 to r = .67 depending on the combination of games used in the assessment ($p < .001$, see Table 2).

TABLE 2:

Convergent Validity of Game and Interview Assessments

COMPETENCY LABEL	N MODEL DEVELOPMENT DATASET*	MULTIPLE R**
GAME		
Cognition	364-647 (depends on game combo)	.51 to .67 (depends on game combo)
Empathy	294	.42
Influence	343	.45
Influence	343	.45
Conscientousness	343	.67
Extraversion	343	.72
Agreeableness	343	.52
Emotional Stability	343	.62
INTERVIEW		
Adaptability	2,867	.58
Communication	31,772	.56
Compassion	1,943	.61
Composure	2,670	.56
Coordination of People & Resources	1,809	.58
Dependability	1,838	.55
Developing Others	2,862	.62
Drive for Results/ Initiative	1,502	.50
Negotiation & Persuasion	2,361	.53
Problem Solving	2,674	.57
Relationship Building	1,915	.57
Safety & Compliance Orientation	2,224	.56
Service Orientation	2,212	.57
Team Orientation	1,913	.50
Willingness to Learn	1,766	.52

Predictive validity of hirevue assessments across industries:

HireVue Interview Assessments are used for selection across industries, professions, and job levels. They demonstrate correlations ranging from .25 to .49 between assessment scores and job-related outcomes. HireVue Custom assessments use algorithms that directly predict a given job-related outcome, specific to the customer. A custom interview assessments may predict measures such as candidate performance, sales results, or turnover. This is particularly successful when a large number of employees perform similar tasks and are evaluated consistently. For example, an assessment used in the Technology Services industry to predict the sales performance of sales representatives provided substantial utility for the organization. The predictive validity (uncorrected) for this assessment is r = .42, which is considered as having the potential to provide a very beneficial business impact (U.S. Dept. of Labor, Employment and Training Administration, 1999). See Table 3 for more examples from different industries. These values are comparable to the predictive validity of structured interviews in the employment context (McDaniel et al., 1994; Schmidt, 2016; Schmidt & Hunter, 1998), but without the resource intensive requirements of interviewers and evaluators of the interviews.

TABLE 3:

Predictive validity of interview assessments in a range of industries

INDUSTRY TYPE	JOB FAMILY ROLES	PERFORMANCE CRITERIA	STUDY TYPE	INITIAL SAMPLE SIZE	AUC VALUE	CORRELATION COEFFICIENT
Transportation	Driver	Safety Behavior	Concurrent	710	.72	**.34****
Hospitality	Call Center Reservation Sales	Sales Performance	Predictive	404	.71	**.25****
Retail	Sales Associate	One Year Retention	Predictive	696	.68	**.29****
Technology Services	Sales Representative	Sales Performance	Predictive	380	.75	**.42****
Airline	Flight Attendant	Hiring Outcomes	Predictive	53,194	.81	**.49****
Education Services	Tutors	Client Ratings of Tutor Quality	Predictive	6,345	.71	**.35****

Notes: AUC values above .60 suggest the model is able to distinguish between two classes fairly well.

Fairness and adverse impact

There are various mathematical and cultural notions of fairness. In the US, according to the Uniform Guidelines on Employee Selection Procedures (1978), any measure used for selection must demonstrate a lack of adverse impact. Adverse impact is present when applicants from one or more protected groups (e.g., gender, ethnicity) are selected at significantly different rates. Ensuring a lack of adverse impact is a core objective in using assessments to achieve a selection process that promotes diversity.

One example that is prevalent in the US pre-hire assessment process is the 4/5ths rule. Specifically, a selection rate for any protected class (e.g., race, ethnicity, age, or gender) which is less than 4/5ths of the rate for the group with the highest passing rate

is considered to indicate adverse impact. With the example of gender, where male candidates are passing at a higher rate, the adverse impact ratio is defined as:

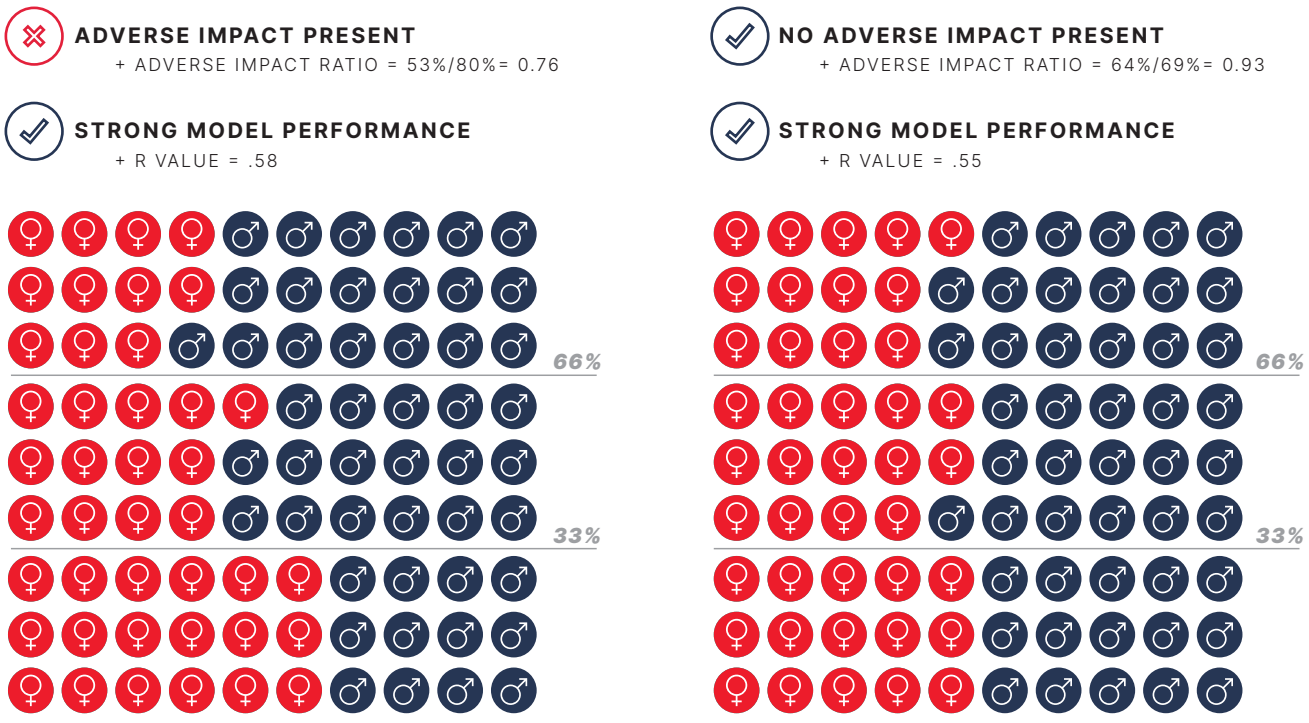
ADVERSE IMPACT RATIO (AIR) = FEMALE PASSING RATE / MALE PASSING RATE

While the 4/5ths Rule presents a quick indicator of substantial disparities in passing rates, the Uniform Guidelines as well as professional standards recommend statistical measures also be used to establish whether adverse impact is present (AERA 1999; SIOP 2003). Statistical tests can help establish whether groups have significantly different scores on the assessments.

Figure 6 shows an assessment model where the initial pass rate of females is significantly lower than that of males, which is below the 4/5ths rule threshold. The HireVue adverse impact mitigation process is applied in order to minimize this adverse impact and establish acceptable selection rates for females compared with males. After the mitigation process the assessment retains its convergent validity (convergent validity correlation of r = .55 after compared with r = .58 before mitigation), while providing comparable passing rates for males and females (AIR = .93 compared with AIR = .76 before mitigation, at an assumed cut off of 60%). An AIR of <.8 indicates selection rates that are less than 4/5th for the minority compared to those of the comparison group. Therefore, the original model with an adverse impact ratio of .76 indicates adverse impact is present and the 4/5ths Rule is violated. After mitigation, the adverse impact ratio is .93, indicating adverse impact is not present.

FIGURE 6:

Example assessment scoring algorithms before and after mitigation



In order to achieve this, a rigorous feature investigation is conducted with the aim of identifying features that have a strong relationship with gender, but little impact on the model performance. These features are identified and removed or de-weighted through an iterative automated procedure, illustrated in Table 4. For the example model shown, cognitive state words make up only .19% of the model performance, but account for 7% in gender differences. **Therefore, the feature is deleted from the model.**

TABLE 4:

Example of feature discounting during adverse impact mitigation

FEATURE	TOTAL FEATURES: 13,335	
	IMPACT	IMPACT ON BIAS
Emotion Words	.2%	0%
~~Cognitive State Words~~	~~.19%~~	~~7%~~ To mitigate adverse impact
Technical Language	1.8%	1.5%
Power Words	.81%	0%
Pronoun Usage	2%	.2%

Features that consistently predict protected classes, or theoretically should not be related to performance at work, are blacklisted and permanently disqualified from our models. For example, the pronunciation of certain words could be correlated with ethnicity. The model is then re-trained without the identified features. All models used in our assessments must pass all our adverse impact tests while maintaining satisfactory model performance (convergent or criterion related validity).

All assessments are tested to ensure that group differences are at acceptable levels while maintaining satisfactory model validity (convergent validity). Table 5 shows the various statistical indicators, as well as 4/5ths Rule values against which all assessments are checked. The Service Orientation assessment shown here has comparable selection rates for different ethnic, age and gender groups. In addition, statistical tests show that there are no significant group differences in assessment results between these groups. Together the 4/5ths Rule and statistical tests show that the assessment should select comparable rates of candidates from different groups. When using assessments for selection, this must be checked regularly to ensure that the respective selection process does not have significant group differences.

TABLE 5:

Statistical and practical adverse impact indices used by HireVue with the example of adverse impact values for the Service Orientation interview assessment.

PROTECTED GROUP	PASSING RATE	ADVERSE IMPACT RATIO	COHEN'S H	PRACTICAL EVIDENCE OF ADVERSE IMPACT?	FISHER'S EXACT TEST	CHI-SQUARED TEST	STATISTICAL EVIDENCE OF ADVERSE IMPACT?
Service Orientation							
Black (n=412)	52.2%	—	—	—	—	—	—
White (n=582)	47.9%	0.92	0.08	**No**	0.20	0.19	**No**
Asian (n=108)	47.2%	0.90	0.10	**No**	0.39	0.36	**No**
Hispanic (n=961)	50.7%	0.97	0.03	**No**	0.64	0.61	**No**

CULTURAL VARIABILITY

The HireVue Science team participates in the wider algorithmic fairness community, attending conferences and keeping up with new research publications. The Uniform Guidelines, professional testing standards, and regulatory agencies across the globe are used as a starting point to set HireVue standards for fairness. Beyond these guidelines, there are various measures of fairness that are considered. In addition to group differences in scoring, differences in model accuracy or false-positive rates by group are examples of other metrics we consider when we select a final algorithm to put into production.

HireVue assessments are frequently used in a global context. This introduces a number of considerations during the development and implementation process. HireVue assessments measure culturally general competencies and traits such as cognitive ability, personality, and job-related skills. Steps are taken in the design process to develop assessment content that takes cultural variance into consideration where possible. For example, cognitive ability game-based assessments make minimal use of language or prior knowledge. HireVue carefully examines the representation of different groups of its training sets to ensure the composition of the training set reflects the population where the algorithm will be used. When gaps are identified, HireVue develops a sampling plan to obtain more data from a target population. It is also important to compare model performance across groups. If cultural differences are identified, either assessment content is modified or algorithms are adjusted. Additionally, it is our practice to regularly monitor for adverse impact and mitigate algorithms for bias if necessary. To ensure that candidates are only compared to others in their country or region, local norming groups are implemented. We localize our assessment content (instructions, test items, and interview questions) using experts in psychometric test translation. All models that rely on verbal behavior are language-specific.

Advantages of interview and game-based assessments

Technology can be leveraged to use data sources other than self-report questionnaires to profile personality and work-relevant competencies. Whether machine learning-based or traditional scoring algorithms are used, in order to develop a psychometric assessment suitable for application in selection or development contexts, one must follow the framework of psychometric theory and practice. This includes the adherence to best practice guidelines and a clear theoretical rationale for the assessment modality used. When developed within these guidelines, interview and game-based assessments promise significant advantages over traditional assessment modalitie

1. HIGH FIDELITY - Having candidates provide responses via interview and game-based assessments more closely approximates the job environment and allows candidates to exhibit behaviors relevant to job performance (e.g., simulating communicating orally with team members through a response, or tracking of information and pattern detection in a dynamic game-based assessment), versus responding to static multiple-choice questions.
2. EFFICIENCY - Interview and game-based assessments sample behavior in a substantially richer and interactive medium than could be achieved with close-ended multiple choice or Likert-type questions typically included in a traditional pre-employment assessment. Accordingly, a broad range of job-relevant competencies can be measured in less than 30 minutes. For example, the HireVue graduate assessment consisting of six interview questions and 3 short game challenges measures eight competencies: cognitive ability, communication, dependability, drive for results, problem solving, team orientation, adaptability, and willingness to learn.
3. CANDIDATE CENTRIC - HireVue was founded on the ability for candidates to share their unique story. This is evidenced by a Candidate Net Promoter Score average above 70 across all of our clients. Interview questions allow the optimal balance between rigorous assessment of candidate behavioral attributes and providing an engaging experience that allows candidates to share their story.
4. MINIMIZES FAKING AND CHEATING - Our interview and game-based assessments minimize susceptibility to candidate faking and cheating attempts and mitigate against test security risks. Unlike a traditional assessment where candidates can view the response options, interview questions are open-ended and game-based assessment levels are procedurally generated to be unique for each candidate. This minimizes the effectiveness of sharing responses with others, selecting the most obvious socially desirable response, or gaining much through practice effects by interacting with similar mediums.

REFERENCES

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C., & Ruble, D. N. (2010). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. Journal of Experimental Social Psychology, 46(1), 166-171.

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bardi, A., Calogero, R. M., & Mullen, B. (2008). A new archival approach to the study of values and value-- Behavior relations: Validation of the value lexicon. Journal of Applied Psychology, 93(3), 483-497.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta–analysis. Personnel psychology, 44(1), 1-26.

Bartram, D. (2005). The Great Eight competencies: a criterion-centric approach to validation. Journal of applied psychology, 90(6), 1185-1203.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. Personality and Social Psychology Review, 23(2), 190-203.

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game.
Computers in Human Behavior, 48, 94-103.

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

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.

Cavazotte, F., Moreno, V., & Hickmann, M. (2012). Effects of leader intelligence, personality and emotional intelligence on transformational leadership and managerial performance. The Leadership Quarterly, 23(3), 443-455.
Chamorro-Premuzic, T., & Arteche, A. (2008). Intellectual competence and academic performance: Preliminary validation of a model. Intelligence, 36(6), 564-573.
Hogan, R., Chamorro-Premuzic, T., & Kaiser, R. B. (2013). Employability and career success: Bridging the gap between theory and reality. Industrial and Organizational Psychology, 6(1), 3-16.

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

Chung, C. K., & Pennebaker, J. W. (2014). Finding values in words: Using natural language to detect regional variations in personal concerns. Geographical Psychology: Exploring the Interaction of Environment and Behavior, 195–216.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. Computers & Education, 59(2), 661-686.

DeRight, J., & Jorgensen, R. S. (2015). I just want my research credit: frequency of suboptimal effort in a non-clinical healthy undergraduate sample. The Clinical Neuropsychologist, 29(1), 101-117.
Mischel W. 1996. From good intentions to willpower. In P.M. Gollwitzer & J.A. Bargh (Eds.), The psychology of action: Linking cognition and motivation to behavior (pp.197-218). Guilford: New York.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dries, N. (2013). The psychology of talent management: A review and research agenda. Human Resource Management Review, 23(4), 272-285.

REFERENCES

Gonzalez-Mulé, E., Mount, M. K., & Oh, I. S. (2014). A meta-analysis of the relationship between general mental ability and non task performance. Journal of Applied Psychology, 99(6), 1222–1243.

Gosling, S. D., Gaddis, S., & Vazire, S. (2007). Personality impressions based on facebook profiles. Icwsm, 7, 1-4.

Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. Journal of personality and social psychology, 93(2), 298-319.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. Journal of Research in Personality, 51, 78-89.

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(3), 621-652.

Kanfer, R., Ackerman, P. L., Murtha, T., & Goff, M. (1995). Personality and intelligence in industrial and organizational psychology. In International handbook of personality and intelligence (pp. 577-602). Springer, Boston, MA.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S.M., Seligman, M. E. (2013). The online social self: An open vocabulary approach to personality. Assessment, 21, 158-169.

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.

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. Personality and Individual Differences, 49(4), 331-336.

Kwon, S., Yeon Choeh, J., & Lee, J. W. (2013). User-personality classification based on the non-verbal cues from spoken conversations. International Journal of Computational Intelligence Systems, 6(4), 739-749.

Lambiotte, R., & Kosinski, M. (2014). Tracking the digital footprints of personality. Proceedings of the IEEE, 102(12), 1934-1939.

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. Personnel Psychology, 67, 241-293.

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

REFERENCES

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. arXiv preprint cs.CL/0205028.

MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. Emotion, 8(4), 540-551.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. Journal of applied psychology, 79(4), 599-616.

McPherson, J., & Burns, N. R. (2005). Assessing the validity of computer-game-like tests of processing speed and working memory. Behavior Research Methods, 40(4), 969-981.

Mischel W. 1996. From good intentions to willpower. In P.M. Gollwitzer & J.A. Bargh (Eds.), The psychology of action: Linking cognition and motivation to behavior (pp.197-218). Guilford: New York.

Miranda, A. T., & Palmer, E. M. (2014). Intrinsic motivation and attentional capture from gamelike features in a visual search task. Behavior Research Methods, 46(1), 159-172.

Mohammadi, G., Origlia, A., Filippone, M. & Vinciarelli, A. (2012). From Speech to Personality: Mapping Voice Quality and Intonation into Personality Differences. MM '12: Proceedings of the 20th ACM international conference on Multimedia, 789-792.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Transactions on Multimedia, 16(4), 1018-1031.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77(6), 1296–1312.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. Personnel Psychology, 61(1), 153-172.

Ployhart, R.E. Staffing in the 21st Century: New Challenges and Strategic Opportunities. J. Manag. 2000, 32,868–897.

REFERENCES

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. Annual review of psychology, 65, 693-717.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological science, 2(4), 313-345.

Sackett, P. R., Gruys, M. L., & Ellingson, J. E. (1998). Ability–personality interactions when predicting job performance. Journal of Applied Psychology, 83(4), 545-556.

Sahu, G. (2019). Multimodal Speech Emotion Recognition and Ambiguity Resolution arXiv:1904.06022

Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. Behavior research methods, 48(4), 1383-1392.

Society for Industrial and Organizational Psychology, Inc. (2003). Principles for the validation and use of personnel selection procedures (4th ed.). College Park, MD: SIOP.

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.

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. Annual Review of Organizational Psychology and Organizational Behavior, 1, 45–65.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The General Inquirer: A Computer Approach to Content Analysis. Cambridge, Mass: M.I.T. Press.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

Ullén, F., Hambrick, D. Z., & Mosing, M. A. (2016). Rethinking expertise: A multifactorial gene–environment interaction model of expert performance. Psychological bulletin, 142(4), 427-446.

Uniform guidelines on employee selection procedures (1978). Federal Register, 43, 38290-3831.

U. S. Department of Labor, Employment and Training Administration. (1999). Testing and Assessment: An Employer's Guide to Good Practices. Retrieved July 21, 2020 from https://wdr.doleta.gov/opr/FULLTEXT/99-testassess.pdf.

Vazire, S., & Gosling, S. D. (2004). e-Perceptions: Personality impressions based on personal websites. Journal of personality and social psychology, 87(1), 123-132.

Ventura, M., & Shute, V. (2013). The validity of a game- based assessment of persistence. Computers in Human Behavior, 29(6), 2568-2572.

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 7(4), 66-87.

Wood, R. T., Griffiths, M. D., Chappell, D., & Davies, M. N. (2004). The structural characteristics of video games: A psycho-structural analysis. CyberPsychology & Behavior, 7(1), 1-10.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. Journal of research in personality, 44(3), 363-373.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036-1040.

FOOTNOTES

1 Hyperparmeters are higher-level model settings such as learning rate and regularization strength that govern how a model learns for a given problem and set of data. For Ridge Regression, for instance, an example of a hyper-parameter fixed prior to model training is alpha. Alpha is used to restrict the magnitude of allowable regression weights of predictors in a model.