HireVue

Industrial-Organizational Psychology

Third-Party Audit of Video- and Game-based Assessments

Executive Summary

October 2020

TABLE OF CONTENTS

1. Context of Audit

1.1. Identity and Credentials of the Auditor

The audit documented within this report was conducted via Landers Workforce Science LLC by its CEO Richard Landers. Dr. Landers earned his Ph.D. in industrial-organizational psychology (IOP) in 2009 and has since then worked as both an academic researcher and as a private consultant at the intersection point of industrial-organizational psychology and computer science. He is Principal Investigator for TNTLAB (Testing New Technologies in Learning, Assessment and Behavior), where his research concerns the use of innovative technologies in the domains of psychometric assessment, employee selection, adult learning, and research methods, and innovative technologies include game-based assessment, gamification, artificial intelligence, unproctored Internet-based testing, mobile devices, virtual reality, and online social media. He is also a Fellow of the Society for Industrial and Organizational Psychology (SIOP).

His academic work focuses upon interdisciplinary contributions at various intersections of psychology and computer science. Much appears in prominent psychology journals, such as *Journal of Applied Psychology, Journal of Business and Psychology, Industrial and Organizational Psychology Perspectives*, and *Psychological Methods,* but he has also published extensively in human-computer interaction outlets, including *Computers in Human Behavior, International Journal of Human-Computer Studies* and *Simulation & Gaming.* He has been funded by the National Science Foundation, the US Army Research Institute, and private industry.

Dr. Landers maintains numerous connections with industry, the media, and IO practitioners. His work has been featured in numerous popular outlets including *Forbes, Engadget, New Scientist, Business Insider,* and *Popular Science*. He is a member of the scientific advisory board of a large human capital and management consulting firm, and he also regularly consults with startups and other organizations seeking advice on the best ways to combine the rigorous assessment techniques of IO psychology with cutting-edge technology.

Dr. Landers' writing and editorial activities are varied. He currently serves as associate editor for both the *International Journal of Selection and Assessment* and *Simulation & Gaming*, and he serves on the editorial boards of five journals, including *Journal of Applied Psychology* and *Journal of Business and Psychology*. He is author of a statistics textbook, *A Step-by-Step Introduction to Statistics for Business,* now in its second edition, and he has developed two edited scholarly volumes: *Social Media in Employee Selection* and the *Cambridge Handbook of Technology and Employee Behavior*.

1.2. Goals of the Audit

Dr. Landers, hereafter referred to as "the auditor," approached this audit with the following overarching goal: to understand HireVue's assessment approaches and standard models for their asynchronous video-based and game-based assessments in terms of their adherence to known best practices in IOP for the development, maintenance, and use of assessments for the purpose of employee hiring and promotion, to include but not limited to standards outlined in the *Uniform Guidelines on Employee Selection Procedures* by the US Equal Employment Opportunity Commission, the US Civil Service Commission, the US Department of Labor, and the US Department of Justice (1978), the Society for Industrial and Organizational Psychology's (2018) *Principles for the Validation and Use of Personnel Selection Procedures*, and the *Standards for Educational and Psychological Testing* by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). The conclusions provided here are only as accurate as the information that was provided to the auditor.

1.3. Steps in the Auditing Process

To conduct this audit, the auditor proceeded through three major stages of report development.

In the first stage, the auditor reviewed HireVue's internal Technical Report and attended meetings in which HireVue provided summaries of its products, product history, philosophy, sales approach, development approach, and other initial information that HireVue deemed relevant to the audit. Total meeting time during this stage was 3.5 hours and review time was 3.0 hours.

In the second stage, the auditor developed a set of targeted questions for response from HireVue based upon what he learned during the first stage and asked for a response from HireVue to each question. This resulted in contributions to a response document by both on-staff IOP teams and data science teams. Additionally, meetings were held between the auditor and various teams to clarify more complex issues. These meetings also spurred the sharing of several additional documents which were used as reference materials. During this stage, the auditor completed all of HireVue's game-based assessments and the video-based assessment from the perspective of a job applicant on a desktop computer. The auditor also reviewed the reviewer/manager interface in an additional meeting. Total meeting time during this stage was 5.5 hours and review time was 6.0 hours. In total, 983 pages of documentation were collected across 19 documents, 8 presentations, and 2 sets of training materials reviewed directly.

In the third stage, the auditor reviewed the final supplied documents and developed a draft of this report. Thus, 18.0 hours of informational meetings, interviews, and review of technical documentation served as the initial input for this audit. Additionally, further questions were asked of HireVue throughout this stage as they arose. An initial version of this report was shared with HireVue for commentary and correction. HireVue then suggested corrections to factual inaccuracies and responded to lingering inquiries. To ensure the integrity of recommendations, the auditor then revised the report using his professional judgment to produce its final version.

2. HireVue's Overall Assessment Strategy

For the purposes of this audit, HireVue generally engages in best practices associated with assessment across its services portfolio. They provide three major services that fall within the IO psychology domain:

- Job analytic services, which includes direct client consultation, primary data collection within client organizations, and expert guidance based upon that consultation
- Provision and scoring of asynchronous video-based interviews
- Provision and scoring of a suite of digital game-based assessments

3. Job Analytic Services

3.1. Overview

HireVue's job analytic strategy is a multi-stage research-based approach toward understanding the competencies required for high performance within client organizations, focusing upon identifying alignment between client needs and HireVue's competency model, which serves as a relatively-infrequently-updated (once a year or less often) guiding framework for HireVue's assessment practices. Broadly speaking, job analysis occurs in three stages: (1) pre-implementation communication and planning with the client organization, (2) data collection managed by HireVue, and (3) reporting and recommendation generation.

3.2. Competency Framework

HireVue focuses its assessment practices around its own internally-developed competency model which at the time of audit consisted of 18 competencies split across four categories, as listed here. At this time, each competency is only assessed using one mode of assessment. For each of the following competencies, game-based assessment mode is indicated with **bold text**. Those that are not bolded are assessed via asynchronous interview.

- Rewarding to Deal With / Work with People, defined as "can form productive and rewarding relationships with others"
  - Communication
  - Coordination of People & Resources
  - Developing Others
  - **Emotional Intelligence**
  - Negotiation & Persuasion
  - Service Orientation
  - Relationship Building
  - Team Orientation
- Able to Do the Job / Work with Information, defined as "have the cognitive ability to effectively process information and data they encounter in the role to drive decisions and action"

    - **Cognitive Ability**
- **Problem Solving**
- Willing to Work Hard / Personality and Work Style, defined as “right level and mix of personality, motivation, and attitudes to meet the people, data, and information demands of the job”
    - Adaptability
    - Composure
    - Compassion
    - Dependability
    - Drive for Results & Initiative
    - Safety & Compliance Orientation
    - Willingness to Learn
- **Big Five Personality Traits**
- Technical Skills / Job Knowledge, defined as “knowledge and skills required for effective performance” (not included in this audit)

These competencies were derived from a variety of sources, including past academic research on employee selection, itself focusing upon Bartram’s Great 8 Competency Framework, as well as by consulting the Department of Labor’s O*NET database, and through their own internal job analyses “conducted over 25 years,” which references the summative experience of HireVue’s Principal IO Psychologist.

3.3. Job Analytic Process

HireVue’s pre-implementation review reflects best practices in job analysis. Several different sources are consulted to triangulate upon relevant competencies and better approach clients with a meaningful research plan. Sources include the following:

- Job descriptions
- Competency information from the organization
- Reports from prior analyses
- Compensation information
- Existing interview materials, guides, and other information
- Descriptions of existing performance evaluation criteria
- Training materials
- Descriptions of current assessment tools

Primary data collection includes the following methods:

- Surveys or 30-minute interviews from 10-15 high performing incumbents in the target job role
- Surveys or 30-minute interviews from 10-15 managers/supervisors of high performing incumbents in the target job role

- 45- to 60-minute focus groups with 2-5 subject matter experiences with expertise in the target job role

Primary data collection using surveys asks participants to rate competencies statements related to HireVue's competency libraries, e.g., Communication is indicated by "Communicates in a clear and convincing manner and ensures mutual understanding. Appropriately tailors message to audience." Focus groups and interviews are generally used as follow up after surveys, to provide additional details or to resolve disagreements.

Surveys include questions on both importance and whether the competency is deemed necessary at time-of-hire. Collecting both pieces of information form a strong basis for the use of the competency as a selection target.

All sources of data are synthesized by HireVue personnel into several specific reports. Competency linkage tables, in particular, are an impressive summary of evidence available and scores obtained for each competency.

3.4. Conclusion

In the opinion of the auditor, the job analytic strategy as conceptualized by HireVue reflects a very high standard of quality and rigor in relation to established standards. In fact, HireVue's practices are generally more extensive than most peer organizations with which the auditor is familiar, creating a strong foundation on which to base subsequent decision-making. The end-result of this rigorous job analytic process is the creation of trustworthy content-related validation evidence on a client-by-client basis. The auditor could identify few significant areas for potential improvement and no evident weaknesses with the job analytic process itself.

4. Asynchronous Video-based Interviewing Platform

4.1. Overview

One of HireVue's assessment methods is its asynchronous (one-way) video-based interview platform, which is broadly speaking based upon Campion et al.'s (1997) *A Review of Structure in the Selection Interview*, which explores the advantages of structured (versus unstructured) interviews and design components for structured interviews that maximize their psychometric reliability, validity, and applicant reactions. Campion et al.'s other recommendations for reporting, administration, and in other areas are also followed, lending credibility to their process and emerging as a strength of HireVue's approach.

HireVue's video interview platform is from an IOP perspective a "highly structured" interview. This label is appropriate, because all questions are completely standardized, there is no opportunity for follow-up questions, and scoring is identical across administrations. This high level of structuring potentially sacrifices applicant reactions in exchange for increased reliability and validity. In past meta-analytic IOP research, structured interviews in general have been demonstrated to be a strong choice for collecting valid data about job candidates. Additionally, interviews themselves are immensely popular in the national and most international markets for job applicants. Thus, HireVue's platform emerges as a compelling choice in the assessment market.

4.2. Evidence Supporting Psychometric Reliability

The traditional psychometric concept of reliability refers to consistency of measurement: scores obtained on a measure of a trait for a person, all else being equal, should be the same regardless of time, place, or other circumstances or contexts of measurement. Put another way: for two people with equal true scores on a trait, or for one person with the same true score at one time and also at a later time, a perfectly reliable measure will produce identical scores on both measurement occasions.

HireVue's video-based interviewing platform does not contain "items" in the traditional sense of a Likert-type or ipsative questionnaire measure. In the modern assessment marketplace, the most common alternative to internal consistency measures of reliability used for such measures is the use of test-retest estimates. HireVue has analyzed the question of test-retest across 181,610 assessments in which data are available across two time points. The delay between these two time points is 37.14 days, with an average Pearson's correlation between time points of 0.72, ranging from 0.51 to 0.82 across competencies (see Table 10, p. 52 of the Technical Validation Report June 2020). Additionally, HireVue calculated test-retest reliability estimates by time lag, finding the highest correlations when tests were two tests of the same competency were completed the same day (r = .89) and the lowest correlations for the longest time period recorded, 90 to 180 days (r = .74), as would be expected of any competency measure that contains a skill component.

Overall, HireVue's evidential basis for claims of the reliability of their video-based interviewing platform are justified given existing collected data and analyses. The reliability of the interview platform meets generally accepted standards.

4.3. Evidence Supporting Psychometric Validity

Although not apparent from its technical manual, in discussion with HireVue staff, in support of content validation evidence, it appears that an IO psychologist with a master's degree or PhD drafts initial questions, which are then reviewed by both a technical writer and the product manager, who may revise to improve readability, fairness, and/or other priorities. HireVue also relies heavily upon a rigorous post-hoc data-driven evaluation strategy.

HireVue uses machine learning to evaluate the relevance of questions to competencies and to maintain its question database. Thus, the evaluation of questions focuses upon examining the success of developed predictive models in the prediction of human ratings of interview content. For example, if an interview question is intended to measure the Composure competency, humans, referred to as *evaluators*, must first rate the videos for Composure, and then mean human ratings are predicted via machine learning. If an algorithm cannot be developed that predicts mean human ratings to a certain desirable level of accuracy, the question would not be retained for use in the measure of Composure. If an algorithm can be developed meeting internal standards, it is then applied across clients.

Due to high standards in a data-driven approach, HireVue will not score any new questions provided by client organizations when conducting a standard interview. This is a strength of HireVue's approach, as it maintains the integrity of scoring and ensures HireVue models are only applied in the standard interview when IO psychologists at HireVue have deemed their use trustworthy.

The standards for evaluator ratings are generally quite high. Frame-of-reference training, generally recognized as the most effective rater training strategy, is used to ensure high-quality rating processes, and as conducted reflects standard practices for such training (Roch et al., 2012). Evaluators are taught background knowledge of the competencies they are rating, recommendations for note-taking, cognitive strategies to adopt, and errors to avoid.

The rating scales used by evaluators are behaviorally anchored rating scales (BARS), themselves reflecting a best practice for behavioral rating scales more broadly. Each of the five BARS anchors, which are Novice, Developing, Intermediate, Advanced, and Expert, is clearly defined with clarifying information provided. Initial training takes around 90 minutes, including calibration exercises in which evaluators make sample ratings which are then compared during a live training exercise. These trainings take place in-person or via live video chat. In general, the evaluator training process reflects best practices regarding the training of raters.

Inter-rater agreement is assessed in a non-standard manner, although the approach itself is not flawed. Specifically, instead of reporting intra-class correlations as would be standard practice, Gower similarity indices are presented instead. Although the report suggests that such indices are likely to be more accurate, nevertheless intra-class correlations are more common and readily interpretable by IO psychologist experts. In the M Scott Myers Award Submission, a more interpretable metric, Krippendorff's alpha, is also presented, which reveals a range of values from .48 to .72, which varies by competency. In general, these statistics are a bit lower than desirable by general standards. For example, in their meta-analysis, McDaniel et al. (1994) report a median job-related structured interview reliability of .84 (median = .89, sd = .15). There is no evidence presented in provided materials suggesting why these numbers might be lower relative to the research literature, potentially revealing a weakness in the approach despite seeming adherence to best practices in data collection. However, an alternative explanation is that these statistics are on different metrics (e.g., if HireVue's assessments concern individual interview questions whereas McDaniel et al.'s include a mixture of overall interview ratings and summative ratings across scales).

HireVue's evidence regarding the nomological net of its focal competencies relies heavily upon the provision of client data, which in turn implies they rely heavily upon client cooperation. Although this is common in the practice of IO psychology, more systematically tracking and organizing these materials is recommended. Studies identified included:

- "Across a number of industries and a diverse set of criteria," area under the receiver operator curve (AUROC) "ranging from .65 to .89 and uncorrected validity coefficients… ranging from .25 to .63, with a mean weighted uncorrected correlation coefficient across 20 studies of .47 (N=83,569)." The criterion measures in these validity studies were simply labeled (e.g., job performance, turnover), thus the specific makeup or prototypicality of the meta-analytic database serving as the basis for this claim was not described in sufficient depth.
- Two studies "with a large not-for-profit organization" discussing a custom algorithm developed to predict scores produced by human evaluators provided by client organizations, where "Multiple R's obtained across Study 1 and Study 2 ranged from .50 to .60. Both of these studies were designed by HireVue and became the basis of their Rater Studies to build the assessment algorithms. Both studies were directed by and conducted with a large not-for-profit company under the direction of a nationally known IO Psychologist. Both studies were reviewed and approved by an independent Advisory panel of globally recognized IO Psychologists (according to Hirevue, 5 of 8 have served as Presidents of the Society for Industrial and Organizational Psychology). Names of these IO Psychologists and the Company were not made available to the Auditor.

Although these studies are somewhat informative regarding the nomological net of HireVue's competencies, three general weaknesses could be targeted for improvement. First, the use of AUROC as the target metric for validation accuracy is limited. Although AUROC is a standard outcome measure for binary outcomes in data science, this is a non-standard approach in the context of IOP and also likely to be unfamiliar even to judges already familiar with employee selection litigation in the event of legal action. Second, there is no simple comparison presented of competency standings. A simple meta-analytic correlation matrix among scored HireVue competencies would provide this additional insight. Third, criterion-related validation evidence appears to be available on a case-by-case basis and should be included in technical manuals.

One study was described that speaks to consequence-related validity. In a study of a retail sales cosmetic role, sales metrics were compared before HireVue's product was implemented in 2017 (N=470) and after it was implemented, in 2019 (N=1084). This report revealed a *d* = .18 increase in average ticket sales and a *d* = .24 increase in average units sold. However, this study did not control for external causes of the increase, which leaves interpretation of this increase susceptible to confounds such as the strengthening economic conditions over the same time period.

4.4. Evidence Regarding Fairness and Test Bias

In the Technical Validation Report (June, 2020) adverse impact pass ratios (using a 34th percentile cut score) and group statistical differences are presented for the interview-based algorithms (see the technical manual's Appendix section for relevant tables). Review of the report and adverse impact calculation tables does not indicate any significant adverse impact concerns at the established pass/fail cut scores. However, it is recommended specific adverse impact studies be conducted with the interview assessments with each customer's applicant data set once assessments are implemented.

Given HireVue's quite literal removal of adverse impact, it is quite unsurprising that its predictions do, in fact, evidence little adverse impact, with similar pass rates across race, gender, and age classes, which are standard concerns in IO practice. Differences associated with religion and color, although also protected by the Civil Rights Act of 1964, are not commonly evaluated in practice and are not examined by HireVue.

HireVue also notes that accommodations are provided for disability when requested, which represents standard practice. The report mentions that accommodations are made to make the testing experience equitable, and if such accommodations cannot be made, an alternative form of assessment is recommended for that candidate. In this way, HireVue avoids making claims as to the validity of its assessment for those with disabilities.

4.5. CONCLUSION

One concern not elsewhere addressed regarding video-based interviews is algorithm transportability. Specifically, HireVue appears to apply algorithms developed using one set of questions within a competency to other questions within the same competency. HireVue has provided data evaluating the success of this in one set of studies discussed in the Technical Validation Report (June 2020, p. 36), describing how in "Study 1," human evaluators were used to generate ratings of interviews, and a predictive model was generated with a correlational validity of .60. When applying that same algorithm to "Study 2", correlations "ranged from .50 to .60." Thus, although HireVue presents some evidence on this point, it could be improved.

Overall, however, HireVue's video-based interviewing platform represents a scientifically well-reasoned approach to providing asynchronous virtual interviews, especially given relatively scant literature informing it. HireVue's algorithmic scoring approach, although innovative, does represent a potential attack vector for those wishing to criticize the platform's validity. This is not necessarily a weakness of HireVue's approach and may in fact be a strength; in a sense, it marries state-of-the-art machine learning research coming out of computer science with IO psychology to produce something better than either could manage alone. However, any such forward-looking approach brings risk that those from both backgrounds will not understand nor appreciate the sacrifices necessary to achieve that.

5. SUITE OF GAME-BASED ASSESSMENTS

5.1. OVERVIEW

One of HireVue's assessment methods is its suite of game-based assessments. There are three classes of assessments within this category:

1. **Cognitive ability games** (Digitspan, Disconumbers, Puzzlepicture, Pathfinder, Symbolspan, Flashback, Colortracks, Numerosity, Singularity, Shapedance)
    - These games are placed into packages to form assessments: Short (2-game), Standard (3-game), and Comprehensive (4-game). At present, there are several combinations of each in use, 2 Comprehensive, 5 Short, and 4 Standard.
    - Most of these games represent gamified versions of classic cognitive assessment tasks, like Raven's Progressive Matrices and the Weschler Adult Intelligence Scales' working memory assessments. Others appear to be original creations, most of which resemble cognitive ability test items or brain teasing type games.
2. **Emotional intelligence games** (Chat-type games, E-motions, Pulse)
    - Chat games are instant messaging simulations, although reminiscent of branched situational judgment tests, in that a limited set of possible responses is provided after each message prompt.
    - E-motions is a gamified version of a classic emotion recognition task.
    - Pulse is a gamified version of a classic go/no-go task used to measure impulsivity.
3. **Personality game** (Portrait)
    - The personality game is a forced-choice (ipsative) Big 5 assessment. Questions are presented with the prompt "I am more like…" with possible responses given in pairs, requiring endorsement of one response from each pair. All response pairs involve comparisons of images; some response pairs also include text (i.e., an adjective label) in addition to images.

In most of its materials, including the Technical Validation Report, HireVue treats the games as a set, which matches HireVue's overall philosophical approach: HireVue's competency matching process is intended to determine for clients which competencies they would best benefit from and which game based assessments should be used to measure those competencies. This process, unless clients opt out of it, is generally managed by HireVue's internal IO psychology team. Thus, within this philosophy, clients do not need a complete presentation of what games are available and what they might measure, as their use of any such games will be mediated through subject matter experts.

The games in the suite vary dramatically in design, development approach, and scientific rigor. As such, the remainder of the discussion of HireVue's game-based assessments will address these differences where such information is known and note where it is not.

5.2. Evidence Supporting Psychometric Reliability

For cognitive ability games, test-retest estimates of uniquely defined test batteries vary from 0.60 to 0.81, depending upon the specific battery tested, which varies from slightly below to slightly above generally accepted standards depending upon the battery. In general, longer batteries with more games are more reliable than shorter batteries, as would be expected.

There is no reliability information available for emotional intelligence games.

In a study of test-retest reliability over two weeks (N=107), Portrait (personality game) revealed test-retest reliability estimates ranging from .55 to .75 with a mean of .63 across Big Five personality traits. Although this mean is somewhat low in comparison to traditional Likert-type assessments (mean .73; Viswesvaran & Ones, 2000), they are in the expected range for the test-retest reliability of such assessments presented as forced choice, dichotomous measures. In general, Portrait meets generally accepted standards for reliability.

5.3. Evidence Supporting Psychometric Validity

HireVue's game-based assessments vary in approach. Although all utilize predictive modeling, much like HireVue's video-based assessments, they make specific claims regarding construct measurement for each game or game battery. Specifically:

1) **Cognitive ability games** are intended to assess cognitive ability as a battery.
2) **Emotional intelligence games** are intended to assess different constructs conceptually related to emotional intelligence: influence, collaboration, empathy, and impulse control.
3) **Personality games** mimic traditional personality tests, with multiple items contained therein intended to assess Big 5 traits.

From an IO perspective, HireVue's suite of games vary by type in terms of appropriate validation. Whereas emotional intelligence games are designed and intended to assess individual targeted traits, cognitive ability tests are used in various combinations as a battery to assess cognitive ability, and the personality game assesses a set of constructs (the Big 5) with extremely well-known properties. As such, much as with reliability, validity questions will be split by game category where appropriate, but the games will in some cases vary quite widely from each other in terms of approach.

5.3.1. Cognitive Ability Games

Cognitive ability games are primarily designed based upon long-standing and well-documented cognition measures appearing in the scientific literature. However, some games are not. For all cognitive ability games, it is clear from both review of the game rules and from experiencing gameplay that each is highly cognitively loaded and likely highly correlated with general cognitive ability. Additionally, some features of the cognitive ability games do not appear literature-driven from the perspective of IO, driven by general game design principles instead.

The primary evidence related to the nomological net evidence for the cognitive ability test batteries are from their own validation. Specifically, scoring models were developed to predict a composite of the International Cognitive Ability Resource (ICAR) and the Cognitive Reflection Test (CRT), and these models achieved a cross-validated r of .51 to .67 from a sample of 328 participants from an online panel according to the Technical Manual. According to the Game Assessments Deep Dive Slide 16, single game convergence ranged from .35 (Pathfinder) to .50 (Digitspan) when building models from individual games. As a comparison, HireVue offered a Cubiks study which found convergence between Cubiks and other commercial cognitive ability products between .219 and .648. This evidence is presented to suggest favorable comparisons of HireVue's games versus existing industry games, in that convergence appears to be at a similar level.

As such, existing convergent evidence presents a strong case that the original unmitigated scoring models of the cognitive ability games appear much like traditional cognitive ability tests. In this sense, it appears that the game-based assessments successfully capture most of the variance in general cognitive ability without capturing the biases traditionally associated with it.

5.3.2. Emotional Intelligence Games

The chat game cluster of emotional intelligence games are situational judgment tests. They are relatively light on game elements, making them closely resemble such tests as traditionally administered. The two major game elements incorporated are narrativization, which seems to make these branched tests, and immersion, by presenting the test as a text messaging exchange.

The second cluster of games are those intended to measure specific emotional intelligence "subcompetencies." E-motions is claimed to measure empathy whereas Pulse is claimed to measure impulse control. Both are modeled closely on existing tasks found in the literature and are well-supported in existing research.

Convergent validity assessment and evidential quality varies by assessment.

E-motion, which is intended to measure empathy through gamified emotion recognition in pictures and videos, was validated against the Geneva Emotion Recognition Test in a study of 186 people from a Prolific panel across a wide range of salary, age, employment status, and age, with an approximately even balance across men and women. Convergence was .41, a Pearson's correlation, suggesting 17% overlap in variances between the two measures. This clearly demonstrates that E-motion and this emotion recognition test do not assess the same construct; however, they do converge at a level as might be expected given their differences. For example, emotion recognition on the Geneva test is done by watching audio-video clips, whereas E-motion is a mix of audio and audio-video. Additionally, the Geneva test examines expression of 14 distinct emotions, whereas E-motion only distinguished between 6. It is also unclear to what degree unreliability of both the Geneva test and E-motion might have attenuated this correlation.

Pulse has not been validated against any existing impulse control measure.

TeamChat and LeaderChat are both gamified situational judgment tests intended to assess influence and collaboration dimensions of HireVue's broader emotional intelligence competency. Both were validated against the Situational Judgment Test of Emotion Management, which was itself designed to assess how people manage others' emotions, adapted to workplace contexts for questions. This test is a traditionally scored SJT in that response choices have subject matter expert derived weights associated with each choice which drive scores. TeamChat correlated .45 with this test score and LeaderChat correlated .36 in a study of 343 Prolific online panelists. As TeamChat and LeaderChat were not designed to assess precisely the same constructs as this test, these estimates indicate that these two games do indeed sample from the same general domain as the test.

5.3.3. Personality Game

As an ipsative text-and-image-based measure of personality, Portrait has some unique content-related validity evidence concerns. Specifically, although all comparisons contain images, some also contain text, and it's unclear if this creates interpretation conflicts for assessees.

As for the collection of the image pairs themselves, the Portrait Technical Manual explains that images were selected to create pairs reflecting trait facets, and facets were selected from prior research on the International Personality Item Pool (IPIP), a public domain and generally accepted personality item framework. Images were selected to reflect either high-low comparisons within a facet or as a forced choice between facets. Although these two types of comparisons are uncommonly used in personality tests, HireVue's reliance upon machine learning obviates the usual challenges associated with mixed scaling procedures.

Refinement of the image selection occurred over the course of two studies. In the first (N=300), images were identified by seeking maximum differences in relation to the questions identified above. In the second study (N=431), some items were dropped and a scoring algorithm was developed using HireVue's standard modeling and debiasing approach. Once predictors were identified this way, items were assigned a value of +1 or -1 depending upon their standing related towards the target trait. Finally, the length was reduced by examining various combinations of traits. In general, except for the use of HireVue's standard algorithmic debiasing, this process reflects the development of a forced choice personality assessment according to current industry norms.

Much as with the other types of games, other scales are provided to examine patterns of convergence and distinction between Portrait scores and other more-established scores of the same constructs. These results are summarized in Table 6 of the Portrait Technical Manual and presented in greater detail in a multi-trait multi-method matrix in Table 7. On initial review, Portrait demonstrates reasonable convergence; scores on the IPIP and Portrait converge by their Pearson's correlations between 0.50 and 0.75 (i.e., 25% to 56% overlap in variance). Although these numbers clearly suggest that the IPIP and Portrait do not measure identical personality constructs, they are in line with typical overlaps observed between commercial and academic personality measures of the same construct. In short, Portrait assesses a version of the Big 5 similarly to how most Big 5 measures assess a version of the Big 5.

One possible exception is Emotional Stability. The concern with this trait is observable when examining Table 7; it is more strongly correlated with the Extraversion scale of the IPIP than with the emotional stability measure from the IPIP, and it is also highly correlated with the Portrait measure of extraversion. This suggests the emotional stability scale may be more of an extraversion scale than an emotional stability scale.

5.4. Evidence Regarding Fairness and Test Bias

Given HireVue's quite literal removal of adverse impact, as expected its predictions evidence little adverse impact, with similar pass rates across race, gender, and age classes, which are standard concerns in IO practice. Differences associated with religion and color, although also protected by the Civil Rights Act of 1964, are not commonly evaluated in practice and are not examined by HireVue.

As related to disability, HireVue notes that accommodations are provided when requested, which represents standard practice. The report mentions that accommodations are made to make the testing experience equitable, and if such accommodations cannot be made, an alternative form of assessment is recommended for that candidate. In addition to this, questions perceived to be related to mental health were avoided in the development of Portrait's emotional stability scale. This reflects a high degree of attention paid to reducing disability-related effects.

Although socioeconomic status is not a protected class, it is advantageous to demonstrate that socioeconomic status does not directly contribute to unfairness or unfairness perceptions. One potential cause of status-covarying effects is device type; specifically, because it is more common for people with lower socioeconomic status to have access to the internet only via a smartphone, assessments that penalize people who use smartphones may be unfair to those of low socioeconomic status. In the case of HireVue's game-based assessments, two studies are reported comparing mobile and desktop users, although it is unclear if they are the same study. In Game Assessment Deep Dive, the only study reported is one consisting of 270 Prolific panelists who completed "games" on both desktop and mobile, counterbalancing order. Although the specifics of these comparisons were not reported in materials available to the auditor, it was reported that "a series of paired t-tests" were conducted, finding a "small advantage" for mobile that was "not statistically significant" and of "small effect size." Assuming these interpretations to be accurate, this would provide reasonable evidence that the test is not biased by device type.

5.5. Conclusion

Overall, the auditor offers the following general conclusions regarding validity evidence.

1. Overall, HireVue has collected a significant amount of evidence to support its claims of reliable and valid testing using its games. Although the quality and type of this evidence varies by game, it generally supports that HireVue is measuring what it claims to be measuring using the tests it has developed. Although there is room for improvement, HireVue's test batteries generally meet industry norms for the rigorous psychometric development of assessments.

2. Evidence collected to support HireVue's tests does not always clearly differentiate between development samples, cross-validation samples, and generalizability samples. For example, it appears that many "concurrent validity" estimates were in fact obtained

from the same samples that were used to develop scoring procedures. In IO psychology, evaluation against holdout samples has not historically been accepted as a technique to establish generalizability due to its vulnerability to sampling assumption violations (see Murphy, 1983). Thus, there is a need to better label sample identities throughout materials.

6. General Conclusions

1. Overall, HireVue has provided compelling evidence of a thoughtfully considered approach to employee selection. Its emphasis on rigorous job analytic methods followed up by IO-managed selection of competencies and assessment tools creates a firm foundation for hiring. The use of algorithmic scoring and bias mitigation, although unconventional, does produce scores that are generally supported by collected validity evidence. In general, HireVue reaches or exceeds industry standards for the creation of high-stakes assessments, and this audit exposed no weaknesses that critically undermine HireVue's approach. There is room for improvement in HireVue's product line, as with all tests, but HireVue's claims are at this time generally justified by the body of evidence they have collected here.

2. Because HireVue's approach is unconventional, a major challenge in this audit was reconsidering what HireVue has done in the terms of IO psychology. One approach that might help people do this in the future is to increase emphasis on collecting validation evidence for HireVue's development process rather than validation evidence for individual tests. If, for example, HireVue can produce and clearly document reliability, validity, and reactions evidence supporting the use of their modeling and bias-mitigation approach for any client and for any construct, this provides support not just for their tests as they currently exist but also for any tests they might construct in the future using the same methods.

7. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee select procedures*. http://uniformguidelines.com/uniguideprint.html

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. https://doi.org/10.1037/0021-9010.79.4.599

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single sample designs. *Personnel Psychology*, *36*(1), 111–118. https://doi.org/10.1111/j.1744-6570.1983.tb00507.x

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures*. https://www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in “big five factors” personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*(2), 224–235. https://doi.org/10.1177/00131640021970475