A European Perspective on Psychometric Measurement Technology

Nigel Guenole

Goldsmiths, University of London

Podium Assessment Systems

Cicek Svensson

Caveon Test Security

Bart Wille

Ghent University

Kristina Aloyan

Moscow School of Management, Skolkovo

Peter Saville

10X Psychology

To appear in 'Technology and Measurement around the Globe (TMAG)' from the International Test Commission (ITC). We are grateful to members of the Psychometrics Forum and the European Association of Selection Researchers for participating in this research.

Abstract

The quality of psychological assessment processes in talent management is influenced by our choices about which measurement technologies to use. Technology with relevance to assessing talent is also advancing at great speed in many domains. These advances include processing power and speed, human computer interaction research, and machine learning and artificial intelligence. Given these rapid developments, it is an appropriate time to pause and take stock of how emerging assessment approaches (e.g., game-based assessment) that leverage these new developments are used, relative to more traditional approaches such as questionnaires and interviews. To achieve this objective, we report here on a survey of European assessment practitioners. We ask about the technology they use for psychological assessment, the constructs they measure with those approaches, and the levels of organisations they are used at. We also asked about how traditional approaches are being enhanced with technology and about practitioner perceptions of the reliability, validity and adverse impact and privacy of their technological choices.

Keywords: psychometric testing, psychological assessment, talent management, technology

A European Perspective on Psychometric Assessment Technology

There have been at least two major technology-driven revolutions in applied psychological measurement in industry that have occurred in living memory. The first involved the shift from paper-based to remote computer-based testing, following the wide-scale availability of the Internet. During this, the first notable wave of technological influence on testing, the field witnessed considerable advances in psychometric models for measuring latent traits. These are now routinely implemented in large-scale computerized adaptive testing programs. Computer-based testing was available prior to the Internet, of course, but the development of the Internet certainly encouraged this testing approach. This first technological revolution in testing generated considerable concern. There was worry that the constructs assessed across paper-based and Internet-based testing methods would not be equivalent, and there was concern that unproctored remote testing would lead to higher levels of cheating.

By and large, these concerns were assuaged with the passing of time. Measurement equivalence analyses of the psychometric test data showed that tests functioned similarly between off-line and on-line formats (Mead & Drasgow, 1993), albeit practitioners often make caveats about some screen formats being inappropriate for some assessments (e.g., spatial reasoning on small screens). Concerns about construct equivalence have also lessened as standardized screen display formats for mobile devices have emerged. Despite the fact that fraudulent activity does occur in certification licensing programs, some research revealed that concerns about cheating in remote testing might be overemphasized, as people on the whole do not cheat on psychological assessments (e.g., Nye, Do, Drasgow, & Fine, 2008). Internet-based testing even spurred innovative approaches to remotely proctor test-taking sessions, so long as privacy legislation can be successfully navigated. Some of these solutions involve detecting screen capture attempts, copy and paste prevention, periodic snapshots of the test

taker for identification matching, detectors that identify when more than one person appears in the camera lens, and video recording of the testing sessions. On the whole, the industry rose to the challenge of making psychological measurements with Internet technology, and numerous sets of best practice guidelines now exist (e.g., Lievens, 2006).

The second revolution is one we are in the midst of currently. It is driven by algorithmic scoring of newly available 'big data' sources, such as unstructured text, chatbot conversations, social media content, sensor data from mobile and other Internet of Things (IoT) devices, and even the process data gathered from game-based testing experiences, to name but a few examples. There is as much, perhaps more, concern about the impact of this new technology on measurement as there was during the time that Internet testing was becoming common. Concerns center on outcomes such as reliability, validity, privacy, and fairness. This time around, however, there is a fundamental difference in the application of technology to psychological measurement. Namely, applied psychologists and technologists are developing psychological measurement approaches that harness data that are unstructured (as opposed to structured) and that oftentimes were not originally intended for psychological measurement.

The unstructured data presents a challenge because our most well-tested psychometric models from classical test theory, factor analysis, and item response theory are for structured data. The secondary usage of data, in turn, creates privacy concerns. For instance, new assessment methods have seen social media content scraped and scored, log data created by players of serious games collected by stealth and analyzed for psychological meaning, and the sensor data from wearable and mobile devices scored to infer standing on psychological traits. We do not expect it would be too contentious to say that technology has not impacted the traditional psychological constructs we hope to measure; perhaps with the exception that temporal perspectives on traditional constructs are now more commonly discussed. Instead,

the impact of technology has been on how measurements are made and how they are reported.

Much has already been written about both waves of technological influence on measurement in regard to how measurements are made and reported. Writings include recent edited volumes by Drasgow (2015), Scott, Bartram & Reynolds, (2017), and Tippins & Adler (2011); special issues with commentaries on the topic of measurement and technology from Morelli, Potosky, Arthur, & Tippins (2017); guidelines from relevant testing bodies such as the International Test Commission (2005); a multitude of guidelines from commercial firms; and even a journal devoted to technology and testing issues, the *Journal of Technology and Testing.* These writings, for the most part, reveal advancements in testing technology and their adoption are an international affair, but with heavy North American influence. This chapter, therefore, will instead provide a European perspective on the state of technology for psychological measurement amongst applied practitioners. Recognizing that the technology advancements themselves are international in nature, we will examine aspects of technology and measurement that can be considered uniquely *European*. That is, we report here on a survey examining the attitudes of European practitioners working in talent acquisition and development, and the attitudes of European workers experiencing these assessment methods during this second wave of technological innovation in measurement.

We apply an industrial-organizational psychology lens, examining attitudes to technology developments in pre-hire situations for talent acquisition (candidate attraction, recruiting, and selection) and post-hire situations for talent development (measuring work attitudes including employee engagement, citizenship behavior, counterproductivity and turnover intentions; skill acquisition at work; and learning and development). We report on the practitioner perspectives in these areas regarding a) the current use of technology for psychological measurement in Europe, b) the adoption of emerging applications of

technology in Europe, and c) the acceptance of technology for psychological measurement in Europe. It is worth noting that the data we analysed was collected in January to April 2021, when the world had acclimatized to new ways of working. We expect that any changes in perceptions of technology for assessment that are attributable to COVID-19 are likely to be reflected in responses.

Method

We used a survey of talent management practitioners in Europe who are using measurement technologies in talent management. We define measurement technology broadly to include common methods people use to take psychological measurements. There are many ways to categorize the measurement technologies that we might have followed, and often, the edges between where one method finishes and another begins are blurred. For instance, questionnaires are a measurement technology frequently administered as part of an assessment center; while assessment centers are a measurement technology in their own right that might include questionnaires. To some degree, boundaries between assessment methods are arbitrary, but there are still distinctions made in practitioners' minds. In this research, we survey our participants about nine different categories of measurement technology based on our experience in industry. The measurement technologies we ask about are *interviews*, *questionnaires*, *assessment and development centers*, *situational judgment tests*, *game-based assessment (GBA), text parsing*, for instance, on resumés and cover letters, *scoring of digital footprints including social media*, and *internet of things technology* such as smart phones.

Participants

The survey we report was a convenience survey that was distributed amongst the personal networks of the authorship team and was also promoted on the LinkedIn professional networking website to The Psychometrics Forum as well as to the European Network of Selection Researchers. We received 229 responses overall, of which 182 were

from Europe. Given the convenience sampling methodology, unsurprisingly, the survey was not balanced across European countries and can certainly not be considered a random sample. The largest number of respondents came from where the authors had strong networks, United Kingdom, Belgium, and Sweden, and also Serbia. The exact breakdown was as follows. Belgium (9%), Croatia (2%), Denmark (1%), Finland (1%), France (3%), Germany (3%), Greece (2%), Hungary (1%), Italy (4%), Netherlands (2%), Northern Ireland (1%), Poland (1%), Romania (4%), Serbia (10%), Spain (6%), Sweden (10%), Switzerland (6%), Turkey (1%), and the UK & Ireland (34%).

Given that the sample is a convenience sample, we do not undertake inferential statistical analyses. Instead, we report descriptive statistics on the prevalence of different forms of measurement technology and beliefs about their use. Importantly, for each method, we asked a screener question that first asked whether a particular method was used, which was asked of all participants. A final question that was also asked of all participants was on whether privacy legislation such as GDPR had impacted their decision to use or not use a particular technology. Questions about specific measurement technologies, however, were only asked of those who answered yes to the question about whether they use the method.

The chapter sections that follow first provide a high-level overview of existing measurement technology research in each area, followed by a discussion of survey results from practitioners who use these assessment methods in Europe. Some of the questions about the assessment methods were common across different assessment methods, while some were unique. The common questions related to topics that are relevant for all measurement methods, principally reliability, validity, adverse impact, candidate experience, and level of seniority at which methods are deployed. Other questions are specific to specific methods, such as the format of the scenario presentation for a situational judgment test, or whether emotions are assessed in automated video interviewing. For each measurement method, we

first present the common questions across methods, followed by results for more specific follow-up questions. Before we proceed, it is worth discussing interpretations of common psychometric terms.

Interpretations of common psychometric terms

Interpretations of reliability. Different forms of reliability are appropriate for different psychometric applications. Internal consistency reliability might be appropriate when gauging whether a set of items is sufficiently homogenous to warrant interpretation of a score as a measure of a construct. In contrast, test-retest reliability might be appropriate for testing the stability of a construct over measurement occasions. Notwithstanding that different reliability forms are important for different applications, test-retest reliability is appropriate for many, if not most, applications in pre-hire recruitment and selection and post-hire applications in talent management. This is because we tend to measure constructs or knowledge that are relatively stable over timeframes relevant to organizational applications, at least in the absence of intervention. For that reason, when we ask practitioners about their perceptions of the reliability of different assessment methods, we refer to test-retest reliability. In terms of reliability benchmarks, we used labels of very reliable (close to .80, or higher), somewhat reliable (close to .70), and unreliable (close to .60, or lower).

Interpretations of validity. A similar case can be made that different forms of validity matter for different applications as was made for reliability. For example, in the context of validity, content validity might be appropriate when assessing the suitability of a measure as an indication of a candidate's domain knowledge, while face validity might be appropriate when gauging whether candidates are likely to have a favorable reaction to an assessment process in a hiring situation. While there are forms of validity with less and more relevance to different applications, predictive validity for job performance is important to many, if not most, applications of assessment in organizations. Therefore, when we referred

to validity in our survey, we were explicit in referring to predictive validity. In terms of predictive validity benchmarks, when we asked practitioners about their views on the validity of different measurement methods, we used labels of very predictive (close to .30 or higher), somewhat predictive (close to .20), and not predictive (close to .10 or lower).

Interpretations of adverse impact. Reliability and validity have been the overriding concerns when evaluating the appropriateness of different selection methods for some time. However, alongside these two critical criteria are others that are commonly considered equally important, and one of these is adverse impact. Adverse impact occurs when the selection rates for a protected group are lower than those of a majority group. Typically, a threshold of 4/5 is used, with adverse impact indicated if a protected group is selected at less than 4/5 of the rate at which a majority group is selected due to the use of a selection tool. Because the 4/5 threshold can be triggered in a sample due to random fluctuation when the threshold is not reached in a population, the 4/5 rule is typically supplemented with a statistical test such as the Z-test for independent proportions, which is equivalent to the chi-test test for independence in a 2 x 2 contingency table. The Z-test is also known as the 2-standard deviation (SD) rule because an absolute value of approximately 2 (1.96) indicates significance at the .05 level (Collins & Morris, 2008). While many organizations that do large volume recruitment undertake adverse impact analyses, general beliefs about the adverse impact of different methods might preclude their use in others. It is therefore essential to assess the attitudes of users regarding the adverse impact of different methods. We asked candidates whether they believed measurement methods would lead to the following adverse impact levels, small (close to Cohen's d of .20 or lower), medium (close to Cohen's d of .50), or high adverse impact (close to Cohen's d of .80 or higher).

Interpretations of candidate experience. In recent times our conversations with users of measurement technology have taken an interesting turn. Reliability and validity

evidence, once the supreme criteria against which the appropriateness of an assessment was judged, are now joined by conversations about the candidate experience. Examples of the sorts of comments that have been made are that 'questionnaires are 20th-century technology, but we need to provide 21st-century experiences' and 'the candidate experience is just as important as the prediction of post-hire performance'. The reasons are many but broadly center around the value of managing brand perceptions that can be created around strong candidate experiences. In a competitive hiring marketplace, strong employer brands attract higher-quality talent. Firms also want to reject certain candidates but still have them as consumers of their products, apply next time a suitable opportunity arises, and recommend the organization to colleagues and friends. Given that there is at least an apocryphal association between newer assessment methods such as games-based assessment and candidate experience, we surveyed users of each assessment method on the views of the candidate experience of each method. We used a Likert scale for this survey question that ranged from 1 representing a very positive candidate experience to a 5 representing a very negative candidate experience.

The method versus construct distinction

We asked about the measurement characteristics for all methods in general rather than in relation to specific constructs. For measurement technologies that are not intended to assess a particular construct, such as SJTs and perhaps games-based assessment, this distinction is not as relevant. People regularly talk of reliability and validity with respect to the method in general because these technologies measure heterogeneous content as opposed to homogenous content. However, for technologies that assess constructs, the reliability and validity of the measure can vary markedly depending, for instance, on whether a cognitive or a non-cognitive construct is assessed.

On the one hand, a general perception of these features of measurement methods might be argued to lack the finer-grained interpretation of assessing perceptions of methods crossed with constructs. On the other hand, other features of measurement technologies impact reliability and validity, such as how carefully the items that are included are chosen, how many items are included in an assessment, and the assessment conditions under which the assessment takes place. We did not separate out the impact of any such features. A reason in favor of this approach is that our experience is that, amongst practitioners, generalized attitudes towards measurement methods are common. Perhaps most importantly, the generalized attitudes approach has the advantage of being consistent with the notion that this chapter is about measurement technologies – each of our question sets refers to the properties of a measurement technology overall. For researchers who are interested in the granular differences in reliability and validity perceptions when the same method is used to assess different constructs, we refer the reader to Hausknecht, Day, & Thomas (2004).

Chapter structure

We have chosen to present our results according to the measurement method or technology instead of other possible ways. Other ways might have included presenting the challenge that the assessment methodology resolved, such as recruitment or learning and development. However, our data collection strategy was to survey users of different measurement technologies about their attitudes and beliefs about a method. We did not ask questions of people about measurement technologies they indicated they did not use. This means that responses to questions about different measurement technologies are not always based on the same samples. Instead, they are based on different subsamples of users. However, our rationale is that we are writing a chapter on measurement technologies used in Europe, and therefore, organizing the results according to the different measurement technologies is appropriate. In the question that is specific to each method (e.g., whether

video, animation, or text was used for SJTs), proportions need not add to 100 because respondents could choose one, some, or all of the responses.

Prevalence of assessment methods in Europe

We asked all practitioners first about their use of each category of assessment methodology. For this question, the sample of respondents answering about each measurement technology was the same and can be compared. The prevalence rates for each of the nine measurement technologies were as follows: interviews (87%), questionnaires (78%), assessment and development centers (51%), situational judgment tests (29%), game-based assessment (GBA) (15%), internet of things technology (2%), text parsing (5%), digital footprint scoring (5%) and chatbots (3%). We show the comparison in figure 1. These results reveal that despite the considerable excitement about these methods and the and marketing power of many of the emerging measurement technology vendors, amongst sample participants in this survey, traction in organizations is limited. Instead, the dominant assessment methodologies used by organizations in our sample are those we know well, interviews, questionnaires, and assessment and development centers. Of the 'newer' methodologies being used, we see situational judgment tests and games featuring most prominently.

The reason for the low adoption rate of these newer technologies is impossible to discern from our survey. Reasons may include the relative lack of maturity in the scientific evidence for the measurement capabilities of these technologies, or perhaps concerns in relation to these measurement methodologies regarding fairness or perhaps privacy, a topic we return to shortly. One thing we note is that our experience is that there is no shortage of firms offering products and services that use technologies like text parsing, IoT, and digital footprint scraping for talent management. We now turn to thoroughly exploring attitudes toward specific assessment methods. In the following sections, results only represent the

opinions of users of each technology. Across methodology comparisons are not advised because different participants used different technologies.

Insert figure 1 about here

Privacy concerns with measurement technologies

A key feature that separates traditional methods of measurement, such as questionnaires and interviews, from emerging technologies is the nature of the data that they analyze. In traditional approaches to measurement, best practice suggests that information is limited to that identified by job analysis as relevant to the job. Moreover, in typical situations, it is analyzed with the examinee's awareness. That is, the information was provided with the intention that it would be assessed. With newer assessment methods, the best practice is still that a job analysis determines the relevance of the information to the assessment process, but the relevance or otherwise of the information a scoring algorithm might use is not so clear cut. Take social media profile information presented on Facebook or Twitter. Personal information here may or may not be relevant for the prediction of job performance, but studies indicate that some employers do look at this information (Levinson, 2010). In most cases, such information will not have been generated by the individual with the awareness that it would be psychologically assessed. Even where the individual does know that they are being assessed, in some approaches like stealth assessment in GBA, the individual does not know what exactly is being assessed. The growing emphasis on privacy globally is shown by the General Data Protection Regulation (GDPR), which specifies how personal data can be processed, which came into force in 2016.

With this background as context, it would be useful to know the extent to which privacy concerns impacted decisions regarding the appropriateness (or inappropriateness) of different measurement technologies. To get at this question, we asked participants whether privacy concerns had impacted their decisions about whether to use specific assessment

methodologies. This question, like the question on the prevalence of different methods, was asked of all participants in the survey. The results are presented in figure 2. We see that overall, concerns with privacy do not appear to be a determining feature with respect to usage or non-usage of measurement technologies. The highest percentage of the sample indicating they did not use a measurement technique for reasons related to privacy was just 16%. However, this was for scoring digital footprints and social media, which is one of the emerging measurement methods that have hallmark characteristics that prompt concern, i.e., data with potentially only fringe relevance to job performance and using data that was not generated by individuals with the express purpose of psychological assessment as part of a job application. In fact, the top five measurement methods that practitioners indicated not using due to privacy concerns were newer assessment technologies involving non-traditional data sources that people generate without full awareness of how these data will be used. These technologies include digital footprint scoring (16%), internet of things technology (14%), resumé parsing (10%), and automated interviewing (6%). Overall, concerns about privacy do exist with respect to assessment methodology, but there appear to be other factors at play governing their use or non-use.

Insert figure 2 about here.

Interviews

Interviews involve assessment of candidates' potential work performance based on their responses to questions from a future employer or manager. Interviews are one of the most popular methods in personnel selection (McDaniel, Whetzel, Schmidt, & Maurer, 1994; Ryan & Ployhart, 2014). They are popular because they permit contact with candidates, which allows observing a candidate's interpersonal behavior, such as communication style, as well as probing technical skills. They also allow examination of credentials identified in other selection methods (e.g., CVs, reference letters). The observation of interpersonal behavior

seems a particularly attractive aspect of interviews. A recent survey showed that competency-based interviews were one of the most used selection methods across surveyed organizations in the UK (CIPD, 2020). Existing meta-analyses also show that, among various possible selection methods, interviews are perceived very favorably by candidates (Anderson, Salgado, & Hülsheger, 2010; Hausknecht et al., 2004).

Interviews can be unstructured, semi-structured or structured, face-to-face or digital, synchronous and asynchronous. In structured interviews, the process is highly standardized, meaning that interviewers follow a pre-defined set of questions and procedures for all candidates applying for the job (Chamorro-Premuzic & Furnham, 2010; Dipboye, 1994). Such standardization allows interviewers to use ratings or scales to compare candidates. In contrast, unstructured interviews do not follow the same structure, and the questions may vary greatly from one candidate to another. Unstructured interviews are often guided by interviewers' judgments on what would be the best way to lead the interview process. Semi-structured interviews assume some pre-defined structure and questions but at the same time give an opportunity to vary the interview process between candidates. The typology of structured/unstructured interviews is broadly used in the organizational literature. However, there is great variation in the ways that structured interviews are defined and measured in existing research (Macan, 2009).

There have been numerous reviews on the topic of employment interview validity, including by Arvey & Campion (1982), Harris (1989), Judge, Cable, & Higgins (2000), Levashina & Campion (2007); Moscoso (2000), and Posthuma, Morgeson, & Campion, (2002). While unstructured interviews offer greater flexibility for interviewers and interviewees, structured interviews are generally considered to be more reliable and valid (Levashina, Hartwell, Morgeson, & Campion, 2014; McDaniel et al., 1994). Salgado & Moscoso (2002) examined what is typically measured by conventional interviews (i.e., those

that primarily focus on applicants' credentials, descriptive information of experience and self-evaluative questions) and behavioral interviews (i.e., those that focused on job-related behaviors and experiences). Behavioral interviews were related to social skills, job knowledge, and experience, as well as situational judgment. Conventional interviews were related to the Big Five personality traits, social skills, job experience, and general mental ability.

Another meta-analysis by Roth & Huffcutt (2013) examined the relationship between employment interviews and cognitive ability. Specifically, they reanalyzed the results of earlier work by Berry, Sackett, & Landers (2007) and reported a corrected correlation of .42 between employment interviews and cognitive ability, while Berry's et al. (2007) meta-analysis reported a corrected correlation of .27. In another meta-analysis of criterion-related validity of employment interviews Huffcutt, Culbertson, & Weyhrauch, (2013) estimated the mean-corrected validity for unstructured interviews was .20 and for highly structured interviews to be .70. Employment interviews have also been scrutinized regarding applicant faking (Law, Bourdage, & O'Neill, 2016; Levashina & Campion, 2007; Melchers, Roulin, & Buehl, 2020). Melchers et al. (2020) highlighted that there is limited evidence on the effects of faking as well as ways that such faking can be detected.

While selection interviews can be conducted in live face-to-face format, rapid technological developments offer innovative ways for interviewing candidates. This includes synchronous video interviews (SVIs) and asynchronous or automated video interviews (AVIs), and use of Artificial Intelligence (AI) (Woods, Ahmed, Nikolaou, Costa, & Anderson, 2020). Synchronous video interviews assume that candidates interact with the interviewer online in real-time. AVIs can be conducted at any time without the presence of an interviewer. In AVIs, candidates generally receive instructions on how to complete the interview and how to record their answers to the interview questions. Due to their

asynchronous nature, AVIs may be more cost and labor effective for organizations and help to screen a larger pool of applicants (Lukacik, Bourdage, & Roulin, 2020; Suen, Chen, & Lu, 2019). However, evidence on AVIs reliability and validity as a selection method is limited (Hickman, Saef, et al., 2021).

There is also growing interest in machine learning algorithms to assess applicants' characteristics (verbal and nonverbal behaviors) from employment interviews (Naim, Tanveer, Gildea, Mohammed, & Hoque, 2015). For example, Hickman, Bosch, et al., (2021) developed machine learning algorithms to predict the Big Five personality traits (both interviewees' self-ratings and interviewers' ratings) from video interviews. The findings showed that interviewers' ratings of candidate personality could be predicted more accurately with language-based algorithms than self-reports on personality. In summary, while the forms and ways of conducting employment interviews change and develop, reflecting the advancements in business and selection practices, interviews remain popular among organizations and practitioners.

European trends with Interviews. Interviews are the most frequently used measurement technology we surveyed about, with 87% of respondents indicating their organizations use the method. Participants are relatively accurate in their perception of the prevalence of interviews, with 99% reporting they are widely or very widely used. The most frequent domains in which interviews are used are depicted in the upper panel of figure 3. This highlights that they are most commonly used for recruitment (89%), followed by career management (50%), performance management (37%), learning and development (34%), culture interventions (21%), and finally, restructuring (21%). The most common constructs assessed with interviews are depicted in figure 3 (middle panel). Interviews are most frequently used to assess motivation and interests (77%), followed by technical competencies (KSAOs) (76%), behavioral competencies (73%), leadership potential (57%), personality

(36%), performance management (25%), and general mental ability (18%). Interviews are used extensively across seniority levels as a measurement technology, as indicated in the lower panel of figure 3. Users reported using them at the graduate (85%), experienced (86%), and executive hiring (82%) levels, suggesting usage is consistent across levels. The candidate experience of interviews was rated as very positive or somewhat positive by 70% of respondents.

Insert figure 3 about here

Amongst our respondents, the majority, 62%, considered interviews somewhat reliable, and 33% considered the assessments unreliable. The majority, 70%, considered interviews somewhat predictive, 16% considered interviews very predictive, and 13% believed they were not very predictive of performance. With respect to adverse impact, the majority, 50%, believed interviews led to moderate adverse impact against protected groups, 17% believed they led to large adverse impact, and 32% believed they led to small levels of adverse impact. Traditional and face-to-face interviews are both quite common, with 81% and 72% using each type of interview, respectively. By a long way, the most common form of interview was synchronized interviewing. Only 13% of respondents that indicated they used interviews reported using asynchronous interviews, compared to 90% that reported using synchronized interviews. It is most common to use semi-structured interviews, with 70% of interview users reporting this format, compared to 56% using fully structured and only 14% using unstructured formats. We asked what was scored in interviews. Results revealed that responses to the questions themselves were scored by 79%, body language was scored by 30%, candidate expressions were scored by 29%, intonation was scored by 24%, and text transcripts were evaluated by just 13% of the sample. By a long way, human ratings were the most common scoring approach, with 71% of the sample reporting doing so. Human

qualitative evaluations were reported by 42% of the sample, and algorithmic approaches for quantitative and qualitative evaluations were reported by 2% and 6% of the sample.

Questionnaires. Questionnaires are among the most pervasive technology approaches for psychological assessment, and with good reason. More than a century of research has shown that they can be used to measure the constructs that are relevant to talent management, such as ability, personality, and interests, in a reliable and valid way. In fact, part of the challenge facing new technological approaches to measurement lies in measuring psychological attributes to reliability and validity standards that were originally established using questionnaires. It is not unreasonable to say that as far as the traditional criteria of reliability and validity go, no other measurement comes close to standardized questionnaires. They can be used to assess right versus wrong scored maximum performance cognitive constructs and rating scale based on typical performance non-cognitive constructs that predict the performance outcomes we care about most in organizational settings, such as task performance (e.g., quality, quantity, timeliness of work outputs) and contextual performance (e.g., citizenship behavior, counterproductivity). All the while, questionnaires demonstrate construct validity, meaning well-designed questionnaire-based measures of psychological constructs assess what we claim they assess, which is critically important for feedback. Despite the flexibility of questionnaires, they are not without criticism. The two strongest criticisms of questionnaires, adverse impact and faking, carry the most weight when questionnaires are used to measure two of the most frequently assessed constructs in industrial psychology, cognitive ability and personality.

European trends with Questionnaires. Standardized questionnaires are the second most frequently used measurement technology, with 78% of respondents indicating their organizations use the method. Participants are relatively accurate in their perception of the prevalence of interviews, with 89% reporting they are widely or very widely used. The most

frequent domains in which questionnaires are used are listed in the upper panel of figure 4. This highlights that they are most frequently used for recruitment (90%), followed by career management (57%), learning and development (53%), performance management (43%), culture interventions (41%), and finally, restructuring (20%). Questionnaires are most frequently used to assess personality (95%), motivation and interests (81%), general mental ability (80%), leadership potential (62%), behavioral competencies (58%), technical KSAOs (47%), and performance (28%) (see figure 4, middle panel)). Questionnaires are used extensively across seniority levels as a measurement technology. Users reported using them at the graduate (84%), experienced (83%), and executive hiring (71%) levels, indicating there is a mild drop off in the use of standardized questionnaires at the most senior levels of hiring. These results are illustrated in the lower panel of figure 4.

Insert figure 4 about here

The reliability of questionnaires is viewed relatively favorably by our sample. Forty-six percent believed they were very reliable, 51% believed they were somewhat reliable, and just 3% believed they were unreliable. The pattern was similar for predictive validity, where 47% said questionnaires were very valid, 51% said they were somewhat valid, and 2% said that questionnaires were not valid. Most of the respondents who reported using questionnaires, 45%, reported that they had low adverse impact, with 31% indicating moderate adverse impact and 9% indicating high adverse impact. There were 15% who reported being unsure. The candidate experience for questionnaires was rated as very good or good by 77% of the sample. Just over half, 52%, of the sample reported using adaptive questionnaires.

The most common administrative approach was to use unproctored questionnaires, which 59% of the sample reported doing. Remote supervision (e.g., via camera technology)

was reported by 25% of the sample, and in-person supervision was used by 23% of the sample. The most common administrative method was for desktop computing, reported by 71% of the sample, followed by mobile-enabled assessment by 48% and paper-based administration for just 14%. Appropriate norms were considered important or very important by 76% of the sample.

Assessment and development centers

Assessment centers (ACDCs) are one of the popular and long-established assessment methods in human resource management (Thornton & Gibbons, 2009). An assessment center (AC) involves a comprehensive evaluation of candidates applying for the job, while a development center (DC) involves the same evaluation in a post-hire situation for applications like succession and leadership development. Typically, ACDCs include different elements (e.g., presentations, interviews, simulation exercises etc.) with the aim to observe and examine candidates' behaviors and performance across situations. These behaviors and performance are usually evaluated by multiple assessors who are trained to evaluate participants by following a standardized assessment process (Lievens, 2009; Thornton & Gibbons, 2009). Applicants are commonly invited to take part in an assessment day during which they complete a series of individual or group-based tasks and exercises observed by the company assessors. According to the International Task Force on Assessment Center Guidelines (2015) (hereinafter The Guidelines, 2015), ACs must include clearly defined behavioral constructs related to the job with a clear link to the assessment center elements. Candidate behaviors in the ACDCs can then be classified based on these behavioral constructs. The Guidelines (2015) also underscore the importance of simulation exercises as the key vehicle for ACDC assessment because they provide opportunities for candidates to display behavioral responses to various work-related situations.

ACDCs are valued by organizations because they provide a rounded view of applicants (Thornton & Rupp, 2006). From the applicants' perspective, ACDCs are perceived as a more face-valid method compared to cognitive ability tests (Macan, Avedon, Paese, & Smith, 1994). At the same time, ACs can be very time, cost and labour consuming and challenging to manage over time (Robertson & Smith, 2001). Due to their complexity, ACs also require thorough design and development to ensure the quality and reliability of assessment process. ACDCs have been extensively researched on the topic of their validity. While the criterion-related validity of dimension ratings is well established (e.g., Arthur, Day, Mcnelly, & Edens, 2003; Meriac, Hoffman, Woehr, & Fleisher, 2008), there have been many debates concerning the construct-related validity of ACs. In his review paper, Lance stated that "ACs measure candidate behavior as it relates to the exercises that are constructed and not the dimensions that are defined for assessors to rate candidate behavior" (p. 92). It is rare for psychometric analysis to show more evidence for dimensions than for task performance in assessment centers (for an exception, see (Guenole, Chernyshenko, Stark, Cockerill, & Drasgow, 2013).

The widespread use of technology has also affected the design and implementation of ACDCs. For instance, effective integration of software may facilitate and automatize some internal processes (e.g., scheduling, briefing etc.). ACDCs can also be delivered virtually through various technological solutions and cover a wider pool of potential applicants (Howland, Rembisz, Wang-Jones, Heise, & Brown, 2015). Additionally, virtual ACs may increase efficiency and speed of assessment since candidates are not required to change locations, and the data can be collected and analyzed with technology. This may be particularly valuable to organizations in light of the recent Covid-19 Pandemic and limited opportunities for travel and in-person assessments of candidates. However, while there are apparent benefits associated with virtual ACs in personnel selection, a point of concern is the

impact of increased automatization on interpersonal interactions and subsequent applicants' reactions. This is an issue that requires further research evidence to back up current industry practices.

European trends with ACDCs. ACDCs are the third most frequently used measurement technology with 51% of respondents indicating their organizations use the method. Participants slightly overestimated the prevalence of ACDCs, with 63% reporting they are widely used. The most frequent domains in which ACDCs are used are listed in the upper panel of figure 5. This highlights that they are most frequently used for recruitment (86%), followed by career management (48%), learning and development (38%), performance management (21%), restructuring (15%), culture interventions (7%). ACDCs are most frequently used to assess behavioral competencies (79%), leadership potential (77%), motives and interests (49%), personality (49%), behavioral competencies (42%), general mental ability (39%), and performance (25%) (see figure 5, middle panel). ACDCs are used reasonably frequently across levels, with just under two-thirds reporting using these in hiring graduates (64%), a similar figure to experienced hires (67%). There was noticeably less use of ACDCs for executive hiring, with just over half of ACDC users applying the method to this form of hiring (52%). These results are illustrated in the lower panel of figure 5.

The retest reliability of ACDCs is perceived as favorable, with 32% indicating they produce very reliable data, 65% indicating they produce somewhat reliable data, and 4% indicating ACDCs produce unreliable data. Their predictive validity is seen very favorably by our sample, with 54% indicating scores are very predictive of performance and 46% indicating ACDC scores are somewhat predictive. Most users, 41%, believed ACDCs led to low adverse impact, while 29% believed the adverse impact was moderate, 11% believed it

was high, and 19% were unsure. The most common format was still in person, 77%, followed by supervised remote ACDCs, 47%, and ACDCs that were remote with unsupervised elements, 13%. The candidate experience of ACDCs was reported as very positive or somewhat positive by 93% of users.

Situational judgment

Situational judgment tests (SJTs) are a measurement methodology that can be used to assess a variety of job-related knowledge, skills, and abilities (Cabrera, McDaniel Cabrera, & Nguyen, 2001; Lievens, Buyse, & Sackett, 2005; Lievens & Sackett, 2007; Motowidlo, Dunnette, & Carter, 1990; Weekley, Hawkes, Guenole, & Ployhart, 2015; Weekley & Jones, 1999). They are usually comprised of a stimulus describing a challenging managerial scenario and require a response of the test taker. The stimulus format can take different presentation forms. Most commonly, text is used, but it is also common to see animations and video because of the impact this has on candidate experiences and research suggesting these formats lead to lower adverse impact (Chan & Schmitt, 1997). The response formats are typically rating scales and rankings or partial rankings. However, it is also possible for the candidate to write a text response, or today record a video or audio response. When candidates are told to indicate what would be most effective, scores tend to correlate more with cognitive abilities, and when asked to indicate what they would do, they tend to correlate more with non-cognitive constructs like personality (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Because the responses have been, in the past, primarily rankings and ratings, SJTs are considered low fidelity simulations. Sackett & Lievens (2008) considered SJTs samples of work behavior rather than signs of future performance because they are highly contextualized representations of on-the-job situations.

SJTs are a relatively popular measurement method for hiring and for development due to their high face validity, moderate criterion-related validity, and low adverse impact against

protected demographic groups (Chan & Schmitt, 1997; Whetzel, McDaniel, & Nguyen, 2008). SJTs are best considered measures of generalized knowledge, skills, and abilities. As indicated by their often low internal consistency reliabilities (Catano, Brochu, & Lamerson, 2012; Kasten & Freund, 2016), they are a poor option for measuring psychological constructs, at least when considering the reliability and convergent and discriminant validity achievable by traditional psychometric standards with questionnaires (Guenole, Chernyshenko, & Weekly, 2017; Guenole, Chernyshenko, Stark, & Drasgow, 2015; Lievens, 2017). Recent attention in SJT design has turned to how to gamify the candidate experience through the use of candidate immersion approaches (e.g., multimedia technology) and giving the candidate control (e.g. choosing the order in which the assessment is completed) (Landers, Auer, & Abraham, 2020) and introducing branching into the designs of SJTs (Reddock, Auer, & Landers, 2020).

European trends with SJTs. SJTs are the fourth most frequently used measurement technology, with 29% of respondents indicating their organizations use the method. Participants slightly overestimated the prevalence of SJTs, with 40% reporting they are widely or very widely used. The most frequent domains in which questionnaires are used are illustrated in the upper panel of figure 6. This highlights that they are most frequently used for recruitment (79%), followed by learning and development (36%), career management (30%), culture interventions (13%), performance management (11%), and restructuring (6%). SJTs are most frequently used to assess behavioral competencies (62%), KSAOs (45%), motives and interests (43%), leadership potential (43%), general mental ability (32%), personality (30%), and performance (19%) (see figure 6, middle panel). SJTs are perceived by users as most applicable for graduate hiring (75%) and experienced hires (72%). One-quarter of SJT users indicated using SJTs for executive hiring (25%), 72% for experienced hiring, and 76% for graduate hiring. SJT use across seniority levels is shown in the lower

panel of figure 6. The candidate experience of SJTs was considered very positive or somewhat positive by 78% of users.

Perceptions of the retest reliability of SJTs were relatively positive, with 21% believing they were very reliable, 66% reporting they are somewhat reliable, and 13% reporting that they are unreliable. Similarly, the perceptions of predictive validity were positive, with 31% reporting SJTs were very predictive, 63% reporting they were somewhat predictive, and just 6% reporting they were not predictive of performance. Only 9% reported that SJTs produced large adverse impact, 63% reported moderate adverse impact, and 63% reported low adverse impact. Most respondents, 82%, reported using text vignettes for SJT scenarios, 30% used animations, and 20% used video footage. In terms of response formats, 85% used multiple-choice or ratings and ranking options, 25% allowed text responses, and small proportions reported allowing video or audio responses (6% and 8%, respectively). One quarter reported using branching in their SJTs.

Game Based Assessment

Game Based Assessment (GBA) and related concepts refer to an emerging approach to assessing psychological attributes that aim to increase the engagement of test-takers. GBA has been referred to by several names, including serious games and gamification (Georgiou, Gouras, & Nikolaou, 2019). According to Fetzer et al. (2017), however, gamification refers to the inclusion of gaming elements – such as interactive problem solving, sensory stimuli, and the use of technology, to name a few examples – in non-game situations, while GBA refers to utilizing game elements to create a game that will not be strictly used for fun. If the definitions of these two sub-elements of the games assessment literature sound closely related, it is unsurprising. In fact, the authors note that there is no hard boundary between where gamification ends and games-based assessment begins.

Fetzer et al. (2017) propose that the primary objective of game-based assessment is to increase the engagement of examinees in an assessment process. Later in their chapter, they discuss a second objective of game-based assessment as being to offer incremental validity in the prediction of job performance, implying that evidence suggests games indeed offer incremental validity over other measures such as traditional general mental ability tests and big five personality questionnaires. However, the evidence to date that this potential has been fulfilled is limited. In a more recent paper, Melchers & Basch (2021) mention being able to find only a single paper that provided evidence of games-based assessment scores predicting job performance. The resolution of the apparent discrepancy appears to be in the broader interpretation of the notion of game based assessment by Fetzer et al. (2017). These authors consider simulations, such as situational judgment tests and assessment centers, as GBAs.

In this section, we take a narrower view of games-based assessment as involving game elements absent in traditional assessments, such as specific goals for the game and ongoing feedback delivered via technology. From this perspective, the view of Melchers & Basch (2021) appears correct. The evidence of the predictive validity of such games is limited. Very few games-based assessments appear in the published literature demonstrating reliability, construct validity, or predictive validity. One point to recognize in this debate is that games-based assessment is a methodology rather than a construct, so what we would hope to see is evidence of reliability and validity for the constructs we know well, such as cognitive ability and personality, but measured with games-based assessment (e.g., Georgiou et al., 2019; Landers, Armstrong, Collmus, Mujcic, & Blaik, 2021).

The challenge for the field of games-based assessment designers is that often games are back-fitted in the sense that there is little or no empirical evidence of the reliability and validity for a given game, and scoring protocols are rarely available either. The article by Melchers & Basch (2021) is an excellent case in point. These authors reported that,

unfortunately, no evidence on reliability or validity was available and that it was not possible to discuss the scoring algorithm with any degree of clarity.

European trends with games. GBA is the fifth most frequently used measurement technology, with 15% of respondents indicating their organizations use the method. Participants overestimated the prevalence of GBAs, with 29% reporting they are widely used. The most frequent domains in which GBAs are used are illustrated in the upper panel of figure 7. This highlights that they are most frequently used for recruitment (79%), career management (39%), learning and development (36%), performance management (11%), culture and engagement (7%), and restructuring (7%). GBA is most frequently used to assess general mental ability (79%), personality (46%), behavioral competencies (32%), interests and motivations (32%), leadership potential (29%), KSAOs (18%), and performance (11%) (see figure 7, middle panel). While GMA technology is used for hiring at all levels, there is a marked drop off the perception of the appropriateness of this form of assessment from graduate to executive hiring. Eighty-nine percent of users reported using GBA for graduate hiring, 54% reported using assessments for experienced hires, and 21% reported using GBAs for executive hires. This is shown in the lower panel of figure 7. Norms were considered important or very important by 78% of the sample.

Perceptions of the test-retest reliability of GBAs were favorable, with 25% indicating that such assessments were very reliable, 61% indicating they were somewhat reliable, and 7% indicating they were unreliable. Perceptions of the predictive validity of GBAs were also favorable, with 36% believing they are very predictive, 46% believing they are somewhat predictive, and just 8% believing they are not predictive. With respect to adverse impact, 11% reported that GBAs produce large adverse impact, 29% reported moderate adverse impact, and 46% reported no adverse impact. The most common type of GBA was specifically designed games for assessment purposes, 89%, while gamified assessments were

slightly less common, 39%. Just under half, 46%, of the sample reported that candidates knew what they were being assessed against compared to 25% that did not know, and 29% reported candidates were sometimes aware.

Emerging technology

Social Media and Digital footprints

Individuals' online activity on social media channels, such as LinkedIn, Twitter, Facebook, and Snapchat, creates digital footprints which can be used by employers in personnel selection. Recruiters and hiring managers report checking online activity as an additional source of information regarding applicants' characteristics and behaviors (e.g., Deschenaux, 2010; Grasz, 2012). While the use of digital footprints appears a popular practice to guide hiring decisions, there are many unanswered questions related to applicants' reactions to such screening, the legality of such practices, as well as validity of this method in employee selection context (Becton, Jack Walker, Bruce Gilstrap, & Schwager, 2019; Woods et al., 2020). Because social media sites may contain both job-relevant and non-relevant information that may be indicative of different demographic backgrounds, there is also a question regarding adverse impact resulting from selection decisions based on social media profiles (Van Iddekinge, Lanivich, Roth, & Junco, 2016; Wade, Roth, Thatcher, & Dinger, 2020; Woods et al., 2020). Industry seems to be ahead of academic research on the topic of digital footprints and the use of social media, with evidence from academic research just getting underway. For instance, Roth, Bobko, Van Iddekinge, & Thatcher (2016) present a research agenda regarding the use of social media for hiring. They outline the need to explore the underlying processes for social media judgments, questions related to constructs and validity of social media assessments, adverse impact and group differences (e.g., based on age, gender, ethnicity), as well as questions related to applicants' reactions.

European trends with digital footprints. Just nine respondents indicated they used social media scraping for assessment. Of those, five reported using the technology in recruitment, two in performance management, two in culture and engagement, and one in each of learning and development and restructuring. Only two participants believed the approach was widely or very widely used. The most common application was the measurement of leadership potential and competency management with three responses; personality, interests, and KSAOs with two respondents; and GMA and performance, each with a single respondent indicating they used the approach for measuring these constructs. Of the nine respondents using digital footprint scoring, three reported using it for graduate hiring, three reported using it for experienced hires, and two reported using it for executive hiring. Because numbers were so small in this category, we do not report attitudes about reliability, validity, and adverse impact.

Chatbots

Assessment solutions can also include chatbots, which are virtual agents capable of communicating with job applicants at various stages of the selection process. Chatbots are exceptionally versatile. They can be used to conduct initial screening interviews, answer applicants' inquiries, to give updates regarding selection stages, and they can also be integrated as part of other assessments tools. Other routine tasks, such as interview scheduling, can also be performed by chatbots, speeding up the selection process (Nawaz & Gomes, 2020). Given the increased focus on hiring and retaining the best talent in organizations, chatbots may be helpful in screening large pools of potential candidates and reducing the screening pressure from hiring managers. Additionally, they may help to avoid interviewer biases at screening stages since chatbots use automatic and standardized procedures. However, although chatbots can be an attractive tool to aid selection processes, there is very little scientific evidence that explores their relevance or effectiveness in the

context of hiring decisions, nor on applicants' attitudes to this method. The gradual shift towards artificial intelligence (AI) based selection currently observed in the industry needs to be supported by sound research that would enable effective integration of chatbots in future HR practices.

European trends with chatbots. Just five respondents reported using chatbots for assessment. Two people reported using chatbots in recruitment, one in career management and one in the area of culture and engagement. One indicated that none of the areas we asked about described their chatbot application, and no other information is available on what they did do with chatbots. One respondent indicated that chatbots were used for assessment in many areas, including performance assessment, competency assessment, motives and interests assessment, and assessment of leadership potential. One respondent indicated they used chatbots to assess GMA only. One indicated they used chatbots to assess motives and interests only. One indicated they used chatbots to assess personality. With respect to seniority levels where chatbots are applied, three indicated using chatbots with graduate hires, two with experienced hires, and one with executive hires. Because numbers were so small in this category we do not report attitudes about reliability, validity, and adverse impact.

Resumé Parsing

Traditionally, resumés were screened by recruiters to assess the key information about candidates; the AI solutions now offer resumé parsing – automatic extraction, analysis, and storage of relevant information from resumés in an organized manner. Given a large number of resumés that hiring managers may generally go through (and not to mention different styles and formats of resumés submitted by applicants), resumé parsing is an efficient solution for analyzing and structuring information based on specific algorithms for future use by managers. It also may reduce the time spent on initial candidate screening and, as a result,

reduce hiring costs in organizations. Like other emerging AI-based solutions in human resources (e.g., chatbots), resumé parsing method has not been extensively covered in organizational research, although it is generally discussed in practitioner-oriented sources (see, for example, Zielinski, 2016) and software providers. Hence, academic literature lags behind companies that set the trend on using the parsing method as part of their AI-based solutions.

European trends with text parsing. Just ten people reported using text parsing for psychological assessment in their organizations. Of these respondents, five reported using parsing in a recruitment context; three in career management; two in learning and development; two in performance management; and one in culture and engagement. In terms of what gets measured with parsing, technical KSAOs are reported as being measured by five respondents, behavioral competencies by five respondents, performance was reported by two respondents, personality was reported by two respondents, and in each case, a single respondent reported measuring GMA, motives and interests, and leadership potential with parsing. The small number of respondents that reported using parsing indicated they used it across all levels of seniority. Three indicated using parsing for graduate hires, three indicated using it for experienced hires, and two reported using the approach for executive hires. Because numbers were so small in this category, we do not report attitudes about reliability, validity, and adverse impact.

Internet of Things (IoT)

The term ‘Internet of Things’ (IoT) has become increasingly popular in the light of technological advancements, including AI and Big Data, and it describes the interconnection between physical objects and devices through digital networks to exchange data. Nowadays, physical devices (smartphones, tablets, virtual assistants etc.) can gather and share large amounts of information about individuals, and integration of these devices into selection

processes may significantly enhance existing human resource practices. With increasing reliance on ‘smart’ objects used for communication, work, and everyday life, human resources specialists are now exploring and testing IoT-based methods to interact with applicants, and further digitalize recruitment and selection processes. While there is a proliferation of IoT devices organizations are rapidly adopting, there are many unexplored questions from both practical and research perspectives. For instance, the issue of when it is appropriate to use such data, how it should be stored, and what permissions are required remain unagreed, let alone psychometric considerations of reliability and validity. The amounts of data that may be generated through smart devices also require different approaches towards data analytics and data processing. ‘Smart’ devices may be vulnerable to various risks (for example, hacking) and misuse of information. To date, there is very limited evidence from organizational research concerning the application of IoT for employee selection. Most of the discussions come from practitioners in popular press outlets (e.g., Silverman, 2013).

European trends with IoT. Just five respondents indicated that they used IoT technology for assessment in talent management. Of these, the approach was used in recruitment, career management, culture and engagement, and performance management. In terms of what was measured with IoT technology, the few respondents who reported using this technology reported using it across a wide variety of areas, including GMA, technical competencies, behavioral competencies, interests and motives, personality and leadership potential, even performance.

Conclusion

The level of technological innovation has never been as high as today and will probably even be higher tomorrow. Without a doubt, many of these evolutions (think Big Data, Virtual Reality, etc.) carry huge potential for business applications, and the field of HR

and personnel assessment, in particular, are currently discovering several of the opportunities that they have to offer. From our survey among assessment professionals across Europe, it appears that this adoption process is still happening at a relatively modest pace, especially as regards the newer technologies. Nevertheless, 'HR Tech' is gaining momentum, and we generally expect to see a growing number of applications being used in different areas of the talent management process, including assessment. It is a great time to be involved in HR, and assessment in particular.

New technologies come with new opportunities, but of course, they also come with new challenges. However, we prefer to frame this positively instead of pointing at the 'dangers' and 'threats' associated with the (early) adoption of technology. What we are seeing is that the challenges that present itself—many of them inevitably related to reliability and validity—require intense collaboration between practitioners and scientists. Let us not be the practitioner who thinks that science is too slow for these types of applications or that scientific concepts (such as reliability and validity) have become obsolete in this era of fast technological evolutions. On the other hand, let us also not be the overly conservative scientist who can only see the 'validity-threats', fostered by an almost trait-like reluctance to explore potentially disrupting trends.

There are few—if any—scientists who have the resources (or know-how) to develop innovative technologies that can be readily applied in real-life assessment settings outside the lab. Similarly, it may be too unrealistic to expect practitioners to patiently and diligently wade through all phases of the scientific validation process. Again, rather than seeing this as a problem, we should stress the opportunities here. For scientists, this represents a great chance to (finally) do research that has real and direct 'applied value.' For practitioners, teaming up with scientists will boost product quality (and offer competitive advantage

directly and indirectly) in a business world where evidence-based practice becomes increasingly important.

References

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant Reactions in Selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*(3), 291–304. doi:10.1111/j.1468-2389.2010.00512.x

Arthur, W., Day, E. A., Mcnelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125–153. doi:10.1111/j.1744-6570.2003.tb00146.x

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, *35*(2), 281–322. doi:10.1111/j.1744-6570.1982.tb02197.x

Becton, J. B., Jack Walker, H., Bruce Gilstrap, J., & Schwager, P. H. (2019, June 13). Social media snooping on job applicants: The effects of unprofessional social media information on recruiter perceptions. *Personnel Review*, p. 88. doi:10.1108/PR-09-2017-0278

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview–cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, *60*(4), 837–874. doi:10.1111/j.1744-6570.2007.00093.x

Cabrera, M. A. M., McDaniel Cabrera, M. A., & Nguyen, N. T. (2001). Situational Judgment Tests: A Review of Practice and Constructs Assessed. *International Journal of Selection and Assessment*, Vol. 9, pp. 103–113. doi:10.1111/1468-2389.00167

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the Reliability of Situational Judgment Tests Used in High-Stakes Situations. *International Journal of Selection and Assessment*, Vol. 20, pp. 333–346. doi:10.1111/j.1468-2389.2012.00604.x

Chamorro-Premuzic, T., & Furnham, A. (2010). *The Psychology of Personnel Selection*. Retrieved from https://play.google.com/store/books/details?id=TsxRkXKJ97EC

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, Vol. 82, pp. 143–159. doi:10.1037/0021-9010.82.1.143

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *The Journal of Applied Psychology*, *93*(2), 463–471. doi:10.1037/0021-9010.93.2.463

Commission, I. T., & Others. (2005). International guidelines on computer-based and Internet delivered testing, version 2005. *Verfügbar Unter: Http://Www. Intestcom. Org/Downloads/ITC% 20Guidelines% 20on% 20Computer*, *20*.

Dipboye, R. L. (1994). Structured and unstructured selection interviews: Beyond the job-fit model. *Research in Personnel and Human Resources Management*, *12*, 79–123. Retrieved from https://www.academia.edu/download/44957092/STRUCTURED_AND_UNSTRUCTURED_SELECTION_IN20160421-2666-opzc8p.pdf

Drasgow, F. (2015). *Technology and testing: Improving educational and psychological measurement*. Retrieved from https://www.taylorfrancis.com/books/mono/10.4324/9781315871493/technology-testing-fritz-drasgow

Fetzer, M., McNamara, J., & Geimer, J. L. (2017). Gamification, serious games and personnel selection. *Pulakos, J. Passmore, & C. Semedo (Eds.), The Wiley Blackwell*

Handbook of the Psychology of Recruitment, Selection and Employee Retention, 293–309. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118972472#page=312

Georgiou, K., Gouras, A., & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International Journal of Selection and Assessment*, *27*(2), 91–103. doi:10.1111/ijsa.12240

Guenole, N., Chernyshenko, O. S., Stark, S., Cockerill, T., & Drasgow, F. (2013). More than a mirage: A large-scale assessment centre with more dimension variance than exercise variance. *Journal of Occupational and Organizational Psychology*, Vol. 86, pp. 5–21. doi:10.1111/j.2044-8325.2012.02063.x

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On Designing Construct Driven Situational Judgment Tests: Some Preliminary Recommendations. *International Journal of Testing*, *17*(3), 234–252. doi:10.1080/15305058.2017.1297817

Guenole, N., Chernyshenko, O., Stark, S., & Drasgow, F. (2015). Are predictions based on situational judgement tests precise enough for feedback in leadership development? *European Journal of Work and Organizational Psychology*, *24*(3), 433–443. doi:10.1080/1359432X.2014.926890

Guidelines, I. T. on A. C., International Taskforce on Assessment Center Guidelines, Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., … Thornton, G. (2015). Guidelines and Ethical Considerations for Assessment Center Operations. *Journal of Management*, Vol. 41, pp. 1244–1273. doi:10.1177/0149206314567780

Harris, M. M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology*, *42*(4), 691–726. doi:10.1111/j.1744-6570.1989.tb00673.x

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

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

Hickman, L., Saef, R., Ng, V., Woo, S. E., Tay, L., & Bosch, N. (2021). Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal*, (1748-8583.12356). doi:10.1111/1748-8583.12356

Howland, A. C., Rembisz, R., Wang-Jones, T. S., Heise, S. R., & Brown, S. (n.d.). Developing a virtual assessment center. *Consulting Psychology Journal: Practice and Research*, *67*(2), 110–126. doi:10.1037/cpb0000034

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

Judge, T. A., Cable, D. M., & Higgins, C. A. (2000). The Employment Interview: A Review of Recent Research and Recommendations for Future Research. *Human Resource Management Review*, *10*(4), 383–406. doi:10.1016/S1053-4822(00)00033-4

Kasten, N., & Freund, P. A. (2016). A Meta-Analytical Multilevel Reliability Generalization of Situational Judgment Tests (SJTs). *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *32*(3), 230–240. doi:10.1027/1015-5759/a000250

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2021). Theory-driven game-based assessment of general cognitive ability: Design theory,

measurement, prediction of performance, and test fairness. *The Journal of Applied Psychology*. doi:10.1037/apl0000954

Landers, R. N., Auer, E. M., & Abraham, J. (2020). Gamifying a situational judgment test with immersion and control game elements: Effects on applicant reactions and construct validity. *Journal of Managerial Psychology*, *35*(4), 225–239. doi:10.1108/JMP-10-2018-0446

Law, S. J., Bourdage, J., & O'Neill, T. A. (2016). To Fake or Not to Fake: Antecedents to Interview Faking, Warning Instructions, and Its Impact on Applicant Reactions. *Frontiers in Psychology*, *7*, 1771. doi:10.3389/fpsyg.2016.01771

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: development and validation of an interview faking behavior scale. *The Journal of Applied Psychology*, *92*(6), 1638–1656. doi:10.1037/0021-9010.92.6.1638

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

Levinson, M. (2010). Social networking ever more critical to job search success. *CIO. Com*.

Lievens, F. (2006). The ITC guidelines on computer-based and internet-delivered testing: Where do we go from here? *International Journal of Testing*, *6*(2), 189–194. doi:10.1207/s15327574ijt0602_7

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*(1), 102–121. doi:10.1080/13594320802058997

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The Operational Validity of a Video-Based Situational Judgment Test for Medical College Admissions: Illustrating the Importance of Matching Predictor and Criterion Construct Domains. *The Journal of Applied Psychology*, *90*(3), 442–452. doi:10.1037/0021-9010.90.3.442

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: issues and strategies with generating alternate forms. *The Journal of Applied Psychology*, *92*(4), 1043–1055. doi:10.1037/0021-9010.92.4.1043

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2020). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, 100789. doi:10.1016/j.hrmr.2020.100789

Macan, T. (2009). The employment interview: A review of current studies and directions for future research. *Human Resource Management Review*, *19*(3), 203–218. doi:10.1016/j.hrmr.2009.03.006

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). THE EFFECTS OF APPLICANTs' REACTIONS TO COGNITIVE ABILITY TESTS AND AN ASSESSMENT CENTER. *Personnel Psychology*, *47*(4), 715–738. doi:10.1111/j.1744-6570.1994.tb01573.x

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, Vol. 86, pp. 730–740. doi:10.1037/0021-9010.86.4.730

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *The Journal of Applied Psychology*, *79*(4), 599–616. doi:10.1037/0021-9010.79.4.599

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458. doi:10.1037/0033-2909.114.3.449

Melchers, K. G., & Basch, J. M. (2021). Fair play? Sex-, age-, and job-related correlates of performance in a computer-based simulation game. *International Journal of Selection and Assessment*, (ijsa.12337). doi:10.1111/ijsa.12337

Melchers, K. G., Roulin, N., & Buehl, A.-K. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment*, *28*(2), 123–142. doi:10.1111/ijsa.12280

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: a meta-analysis of the incremental criterion-related validity of dimension ratings. *The Journal of Applied Psychology*, *93*(5), 1042–1052. doi:10.1037/0021-9010.93.5.1042

Morelli, N., Potosky, D., Arthur, W., Jr, & Tippins, N. (2017). A call for conceptual models of technology in I-O psychology: An example from technology-based talent assessment. *Industrial and Organizational Psychology*, *10*(4), 634–653. doi:10.1017/iop.2017.70

Moscoso, S. (2000). Selection interview: A review of validity evidence, adverse impact and applicant reactions. *International Journal of Selection and Assessment*, *8*(4), 237–247. doi:10.1111/1468-2389.00153

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *The Journal of Applied Psychology*, *75*(6), 640–647. doi:10.1037/0021-9010.75.6.640

Naim, I., Tanveer, M. I., Gildea, D., Mohammed, & Hoque. (2015). Automated analysis and prediction of job interview performance. Retrieved from http://arxiv.org/abs/1504.03425

Nawaz, N., & Gomes, A. M. (2020). *Artificial Intelligence Chatbots are New Recruiters*. doi:10.2139/ssrn.3521915

Nye, C. D., Do, B.-R., Drasgow, F., & Fine, S. (2008). Two-Step Testing in Employee Selection: Is score inflation a problem? *International Journal of Selection and Assessment*, Vol. 16, pp. 112–120. doi:10.1111/j.1468-2389.2008.00416.x

Posthuma, R. A., Morgeson, F. P., & Campion, M. A. (2002). Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Personnel Psychology*, *55*(1), 1–81. doi:10.1111/j.1744-6570.2002.tb00103.x

Reddock, C. M., Auer, E. M., & Landers, R. N. (2020). A theory of branched situational judgment tests and their applicant reactions. *Journal of Managerial Psychology*, *35*(4), 255–270. doi:10.1108/JMP-10-2018-0434

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology*, *74*(4), 441–472. doi:10.1348/096317901167479

Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social Media in Employee-Selection-Related Decisions: A Research Agenda for Uncharted Territory. *Journal of Management*, *42*(1), 269–298. doi:10.1177/0149206313503018

Roth, P. L., & Huffcutt, A. I. (2013). A Meta-Analysis of Interviews and Cognitive Ability. *Journal of Personnel Psychology*, *12*(4), 157–169. doi:10.1027/1866-5888/a000091

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology*, *65*, 693–717. doi:10.1146/annurev-psych-010213-115134

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419–450. doi:10.1146/annurev.psych.59.103006.093716

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology*, *11*(3), 299–324. doi:10.1080/13594320244000184

Scott, J. C., Bartram, D., & Reynolds, D. H. (2017). *Next Generation Technology-Enhanced Assessment: Global Perspectives on Occupational and Workplace Testing*. Retrieved from https://play.google.com/store/books/details?id=2c9CDwAAQBAJ

Suen, H.-Y., Chen, M. Y.-C., & Lu, S.-H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93–101. doi:10.1016/j.chb.2019.04.012

Thornton, G. C., & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review*, *19*(3), 169–187. doi:10.1016/j.hrmr.2009.02.002

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment Centers in Human Resource Management: Strategies for Prediction, Diagnosis, and Development*. Retrieved from https://play.google.com/store/books/details?id=okt4AgAAQBAJ

Tippins, N. T., & Adler, S. (2011). *Technology-Enhanced Assessment of Talent*. Retrieved from https://play.google.com/store/books/details?id=TbVgLAXYTewC

Van Iddekinge, C. H., Lanivich, S. E., Roth, P. L., & Junco, E. (2016). Social Media for Selection? Validity and Adverse Impact Potential of a Facebook-Based Assessment. *Journal of Management*, *42*(7), 1811–1835. doi:10.1177/0149206313515524

Wade, J. T., Roth, P. L., Thatcher, J. B., & Dinger, M. (2020). Social media and selection: Political issue similarity, liking, and the moderating effect of social media platform. *The Mississippi Quarterly*, *44*(3), 1301–1357. doi:10.25300/misq/2020/14119

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 295–322. doi:10.1146/annurev-orgpsych-032414-111304

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*(3), 679–700. doi:10.1111/j.1744-6570.1999.tb00176.x

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup Differences in Situational Judgment Test Performance: A Meta-Analysis. *Human Performance*, *21*(3), 291–309. doi:10.1080/08959280802137820

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

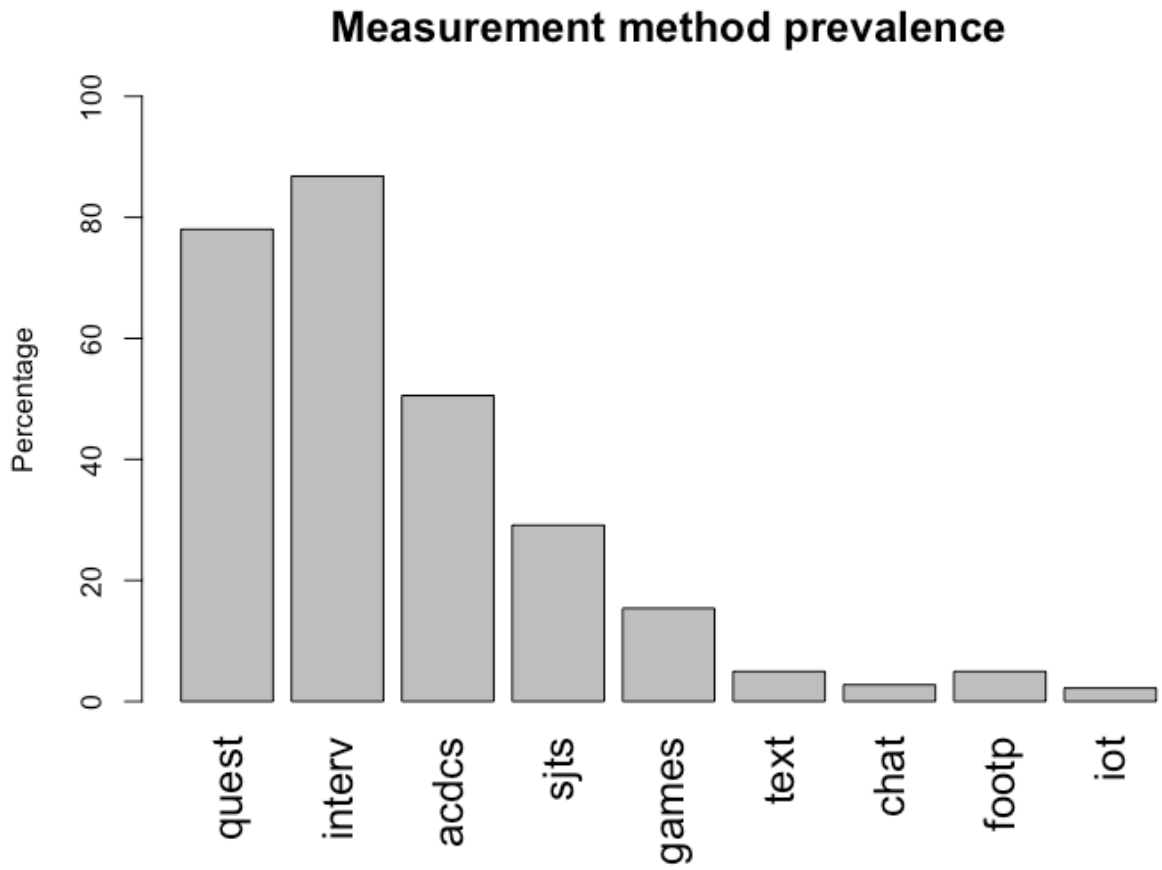


Figure 1. Prevalence of different measurement methods used by European Survey Respondents. N=182. quest=questionnaires; interv=interviews; acdcs = assessment and development centres; games = game-based assessment; text = text parsing e.g., resumés; chat = chatbots; footp=digital footprint scraping; iot = internet of things assessment technology.

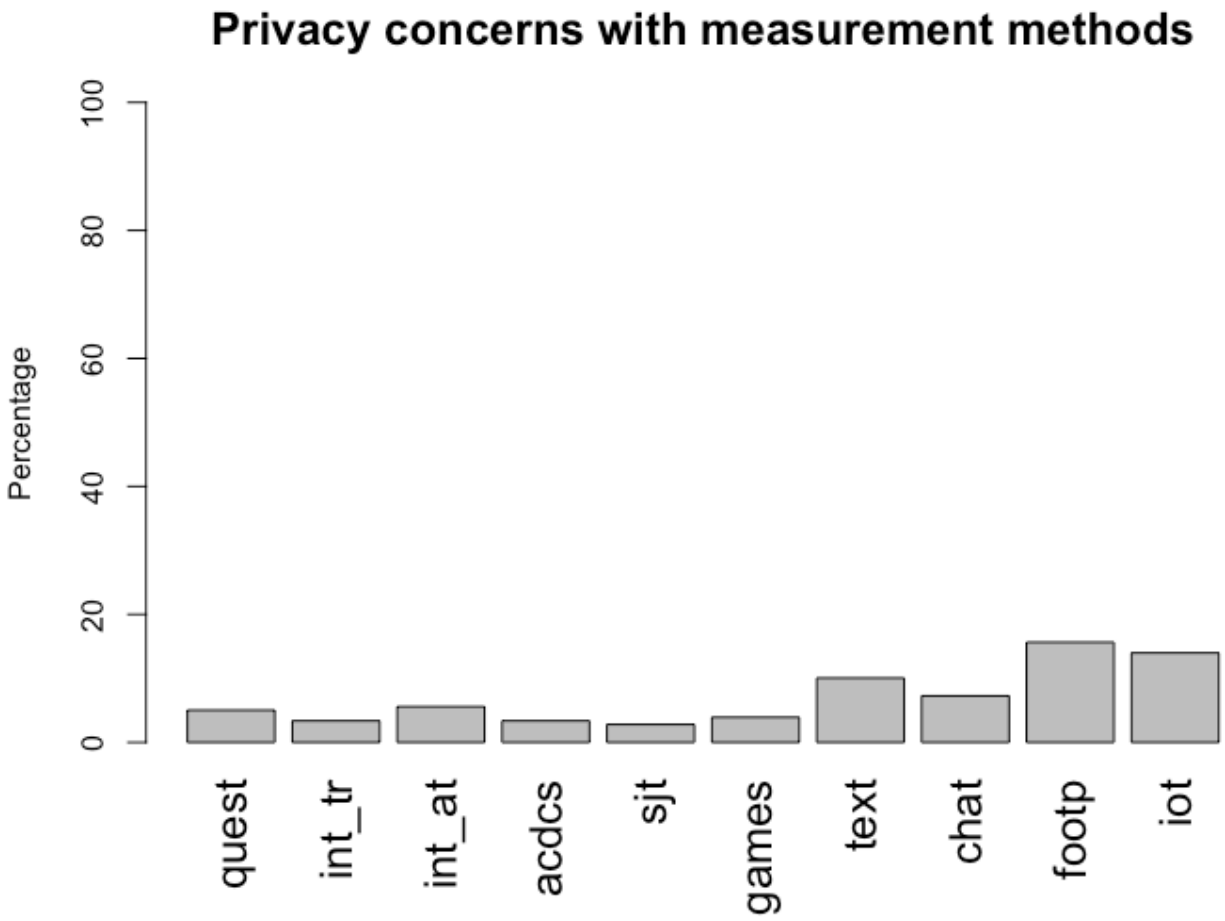


Figure 2. Privacy concerns with measurement technologies. N=182. quest=questionnaires; int_trad=traditional interviews; interv_aut = automated interviews; acdcs = assessment and development centres; games = game based assessment; text = text parsing e.g. resumés; chat = chatbots; footp=digital footprint scraping; iot = internet of things assessment technology.

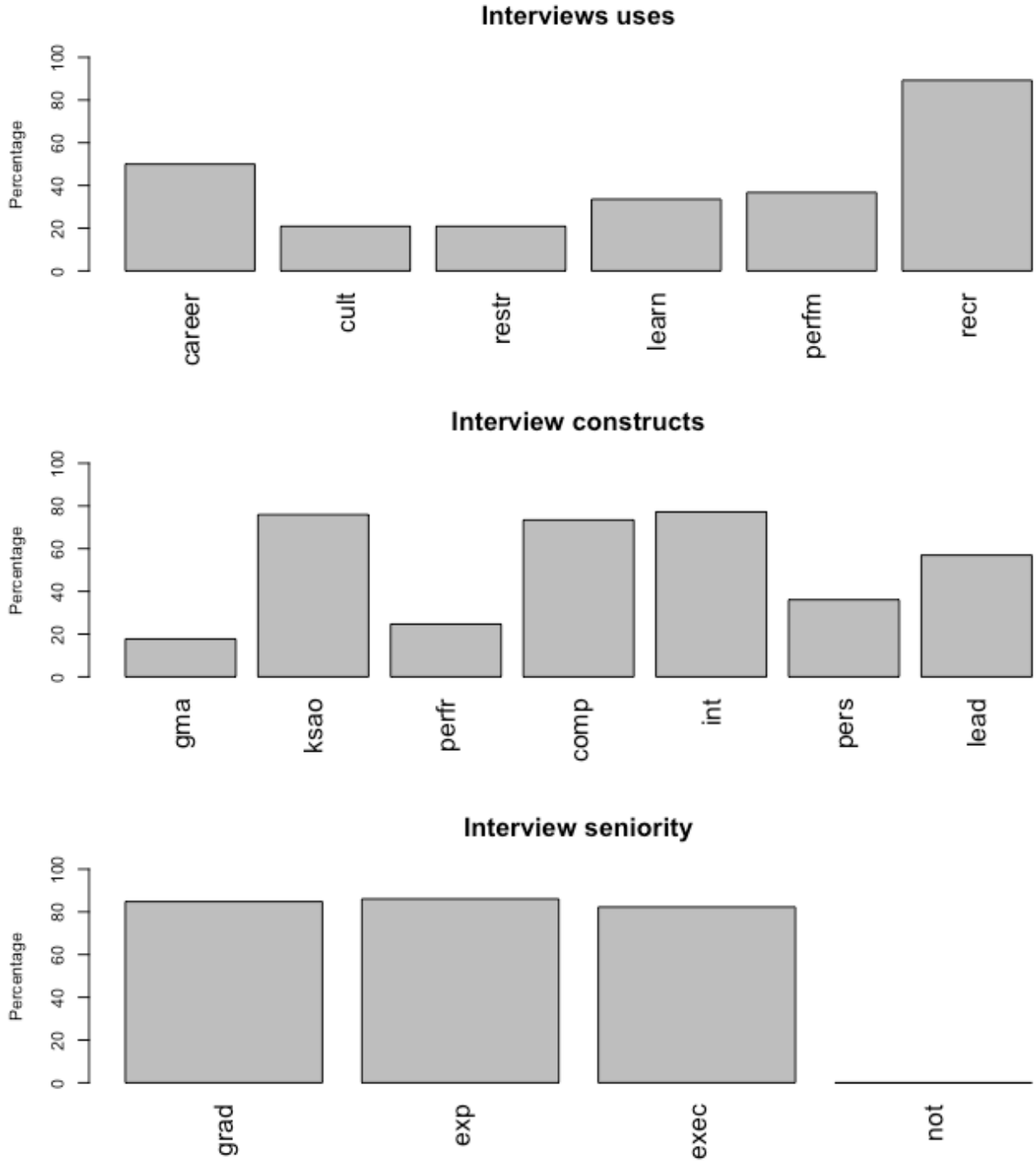
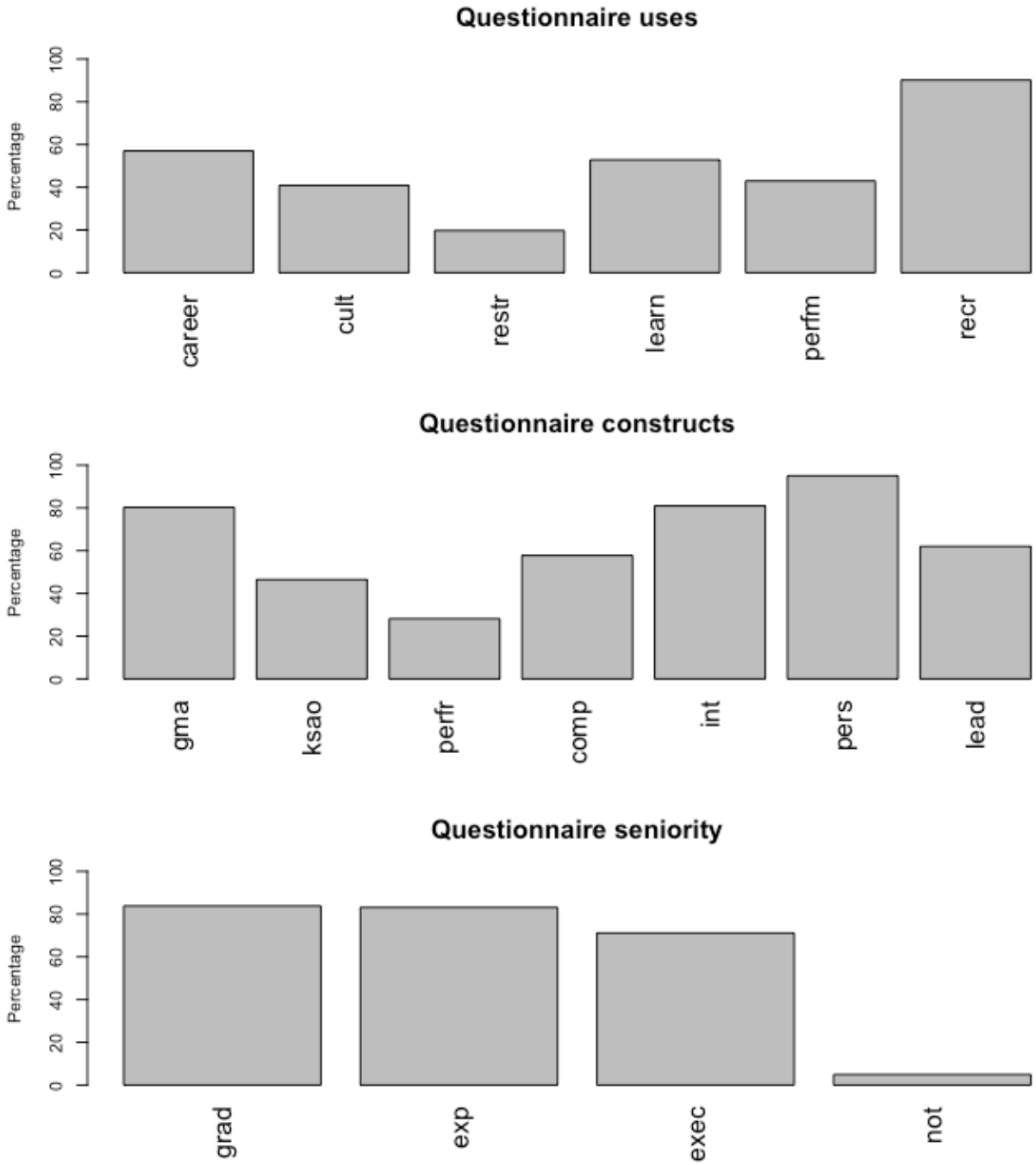


Figure 3. Interviews. *Upper panel.* Use of interviews in common assessment contexts. N=158. Career = career development; cult = organizational culture; restr=restructuring; learn=learning and development; perfm = performance management; recr=recruitment. *Middle panel.* Constructs measured with interviews. N=158. gma=general mental ability; ksaos = general job knowledge; perfr = job performance; comp= competencies; int = motives, values and interests; pers=personality; lead = leadership potential. *Lower panel.* Use of interviews at different levels of seniority. N=158. grad=graduate; exp=experienced hire; exec =executive; not = not used for selection.



igure 4. Questionnaires. *Upper panel.* Use of questionnaires in common assessment contexts. N=142. Career = career development; cult = organizational culture; restr=restructuring; learn=learning and development; perfm = performance management; recr=recruitment. *Middle panel.* Constructs measured with questionnaires. N=142. gma=general mental ability; ksaos = general job knowledge; perfr = job performance; comp= competencies; int = motives, values and interests; pers=personality; lead = leadership potential. *Lower panel.* Use of questionnaires at different levels of seniority. N=142. grad=graduate; exp=experienced hire; exec =executive; not = not used for selection.

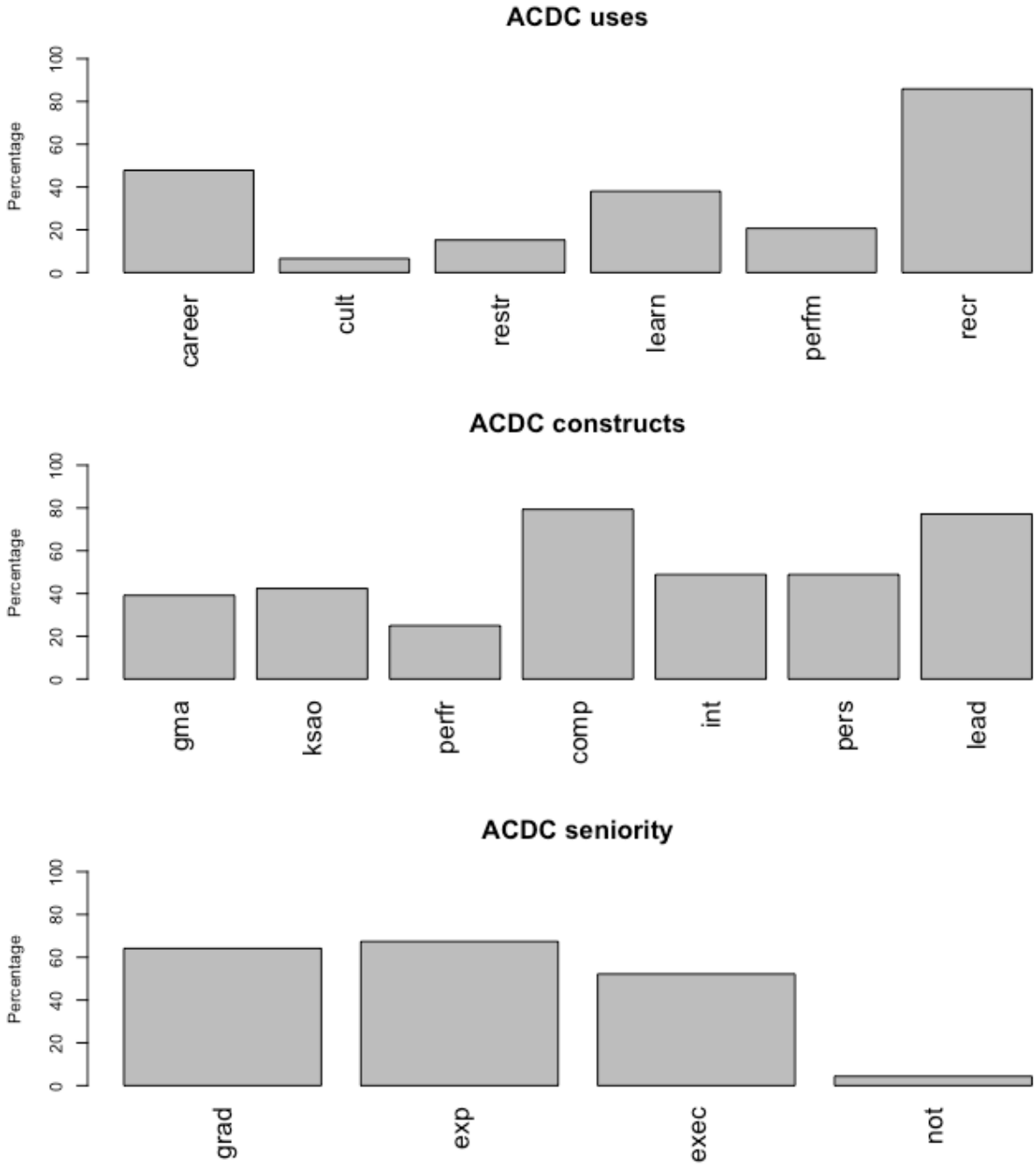


Figure 5. ACDCs. *Upper panel.* Use of ACDCs in common assessment contexts. N=92. Career = career development; cult = organizational culture; restr=restructuring; learn=learning and development; perfm = performance management; recr=recruitment. *Middle panel.* Constructs measured with ACDCs. N=92. gma=general mental ability; ksaos = general job knowledge; perfr = job performance; comp= competencies; int = motives, values and interests; pers=personality; lead = leadership potential. *Lower panel.* Use of ACDCs at different levels of seniority. N=92. grad=graduate; exp=experienced hire; exec =executive; not = not used for selection.

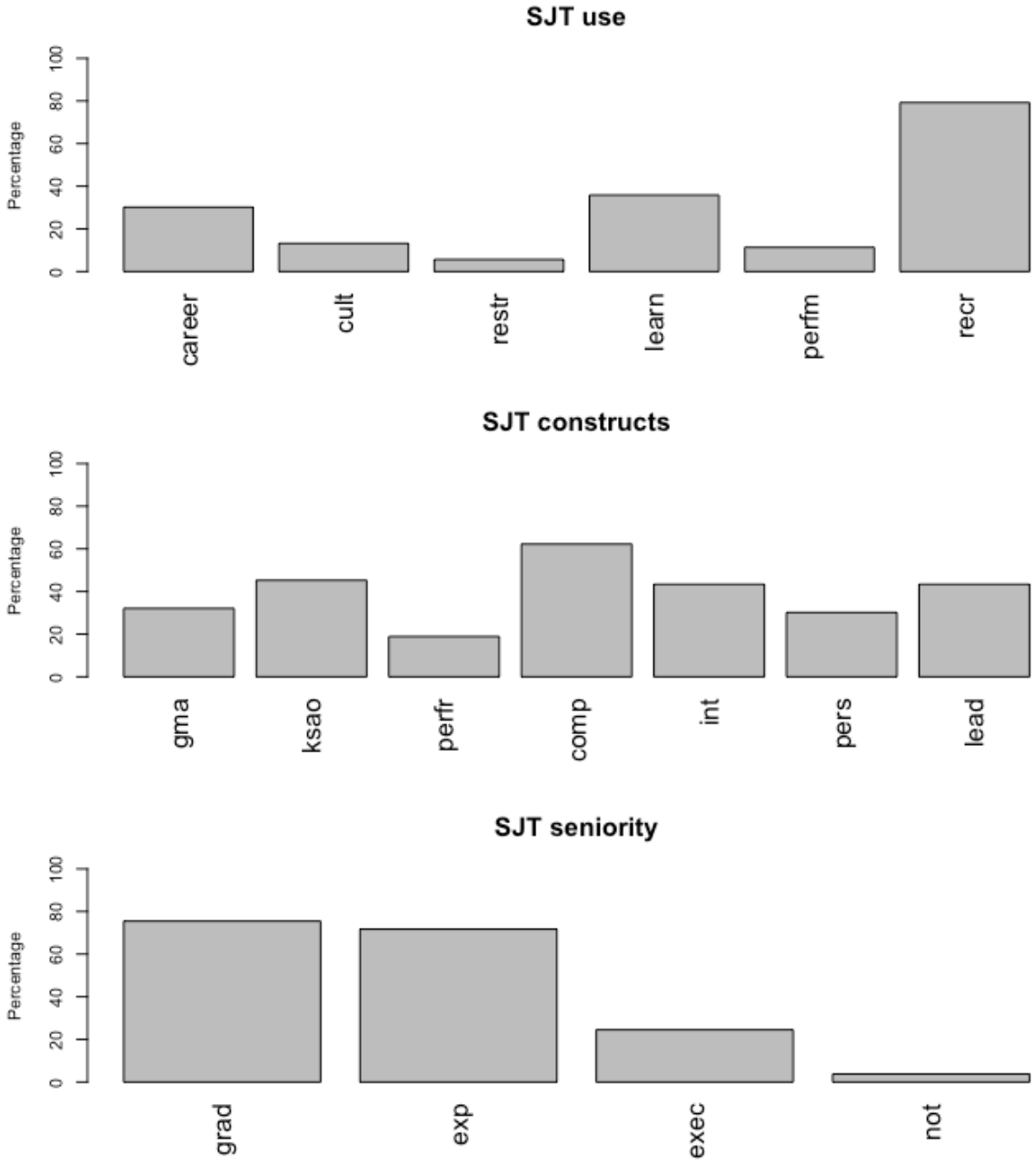


Figure 6. *Upper panel.* Use of SJTS in common assessment contexts. N=53. Career = career development; cult = organizational culture; restr=restructuring; learn=learning and development; perfm = performance management; recr=recruitment. *Middle panel.* Constructs measured with SJTs. N=53 gma=general mental ability; ksaos = general job knowledge; perfr = job performance; comp= competencies; int = motives, values and interests; pers=personality; lead = leadership potential. *Lower panel*. Use of SJTS at different levels of seniority. N=53. grad=graduate; exp=experienced hire; exec =executive; not = not used for selection.

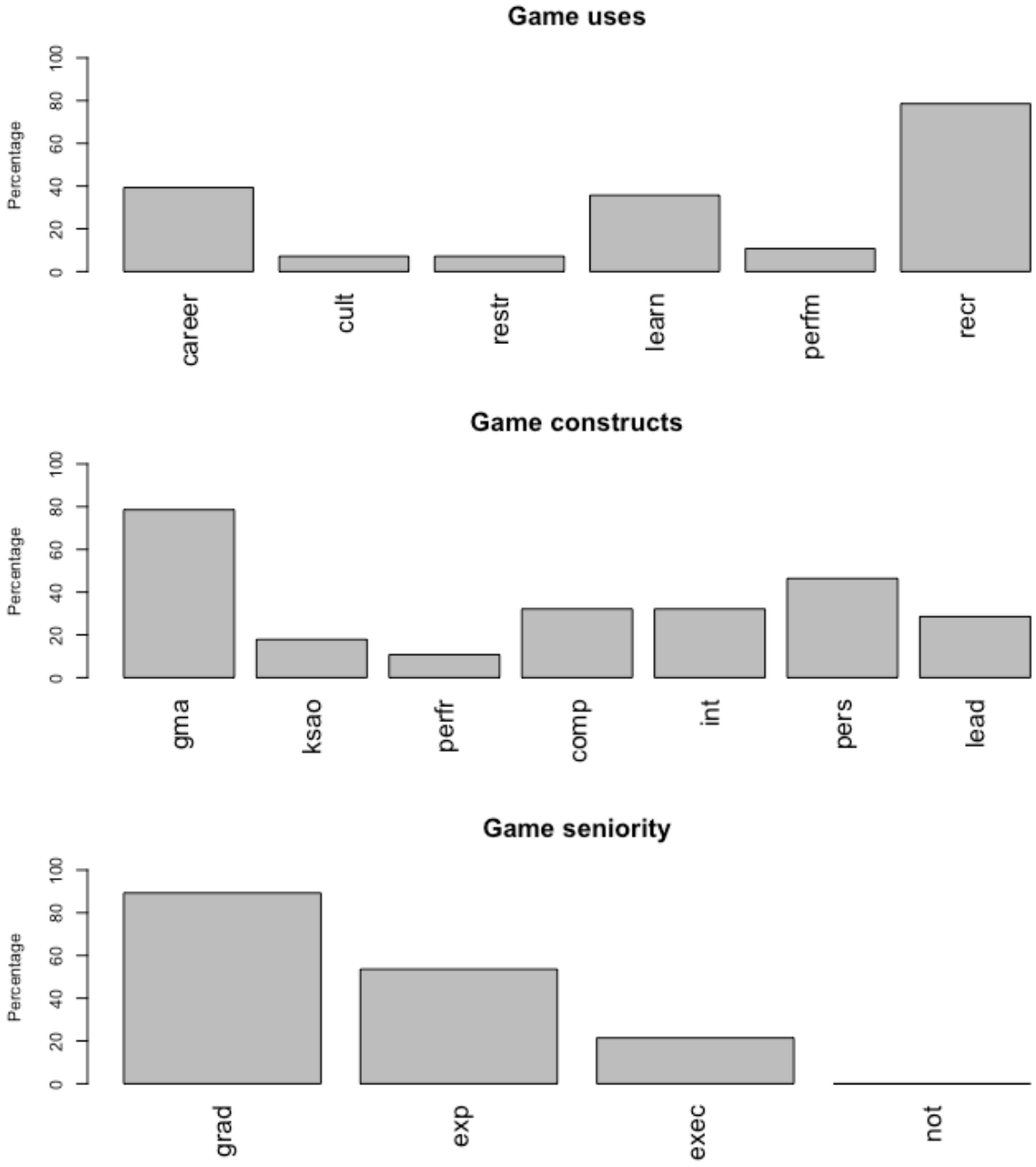


Figure 7. *Upper panel.* Use of GBA in common assessment contexts. N=28. Career = career development; cult = organizational culture; restr=restructuring; learn=learning and development; perfm = performance management; recr=recruitment. *Middle panel.* Constructs measured with GBA. N=28. gma=general mental ability; ksaos = general job knowledge; perfr = job performance; comp= competencies; int = motives, values and interests; pers=personality; lead = leadership potential. *Lower panel.* Use of GBA at different levels of seniority. N=28. grad=graduate; exp=experienced hire; exec =executive; not = not used for selection.