Human Resource Management

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Influence of Proctored Remote Versus Onsite Assessment on Candidate Scores, Assessment Types, Subgroup Differences, and Fairness Reactions

Emily D. Campion[1] | Michael A. Campion[2] | Nicole Strah[3]

[1]University of Iowa Tippie College of Business, Iowa City, Iowa, USA | [2]Purdue University Daniels School of Business, West Lafayette, Indiana, USA | [3]University of North Carolina at Charlotte, Charlotte, North Carolina, USA

Correspondence: Emily D. Campion (emily-campion@uiowa.edu)

Received: 16 July 2024 | **Revised:** 13 February 2025 | **Accepted:** 17 February 2025

Funding: The authors received no specific funding for this work.

Keywords: gender and racioethnic subgroup differences | hiring | proctored and unproctored testing | remote assessments | remote testing | testing candidate fairness reactions

ABSTRACT

As more organizations move to remote hiring assessments, important questions emerge as to the effects on scores, racioethnic, and gender subgroup differences, and candidate reactions. We compare scores of candidates assessed remotely under proctored conditions ($N = 902$) versus onsite ($N = 891$) in an actual selection context in the same organization, in the same time period, and on the same cognitive ability tests, case exercises, and structured interviews. Controlling for job, there were no differences for cognitive ability tests or case exercises in the remote environment, but higher scores for structured interviews, leading to a slightly higher total score for all assessments combined and a 5% increase in the overall passing rate. Within groups, Hispanic or Latino candidates performed better on the remote cognitive ability test compared with Hispanic or Latino candidates onsite, while Asian candidates performed better remotely for the case exercise. All subgroups performed better on the remote structured interview compared with their onsite counterparts. No between-group differences emerged by racioethnicity, but women outperformed men on the remote cognitive ability test compared to onsite. Candidate fairness reactions did not differ by test environment for any assessments or subgroups. We conclude that: (1) remote proctored assessments will not create lower overall passing rates (i.e., fewer candidates for hire); (2) differences in remote assessment scores may depend on the type of assessment, with the greatest positive differences for structured interviews; (3) remote assessments do not disadvantage racioethnic minority candidates or candidates overall; and (4) remote assessments do not reduce candidate fairness reactions.

1 | Introduction

Remote assessments have become common in selection systems due primarily to the speed, access, and cost-saving advantages for both the organization and the candidates (e.g., Jones and Cunningham 2023; Mooney 2002). Although the use of remote assessments began about 20 years ago, the adoption of remote assessments rapidly increased due to COVID-19, which limited face-to-face contact, substantially reduced onsite testing, and impacted hiring in many organizations. Reliance on remote work functions, such as remote hiring, has remained for many organizations long after the health-related concerns of COVID have eased (Maurer 2021; Parket et al. 2022). While quickly evolving technology has made remote assessments of job candidates possible for organizations seeking to assess a variety of attributes for a diverse array of jobs, empirical work has lagged in aiding researchers and practitioners in understanding the pros and cons of remote assessments (cf. Tippins 2015). This lack of

research is particularly salient for how different *proctored* remote assessments influence selection outcomes (e.g., scores and passing rates, subgroup differences, candidate fairness perceptions). We seek to fill this gap by examining these outcomes of proctored remote versus traditional onsite assessments.

The many well-recognized concerns with *unproctored* (e.g., Tippins et al. 2006; 2009; 2015) may be reduced by proctoring. Modern technology has simultaneously enabled proctored internet testing by monitoring with authentication, lockdown browsers, AI to detect possible cheating, and other methods (Nigam et al. 2021), as well as direct monitoring by testing administrators via computer cameras, as in the current study. Even with proctoring, organizations considering implementing remote hiring assessments have at least four major remaining concerns. First, one critical issue is whether the scores and passing rates will differ from onsite assessments. Differences in proctored scores may indicate differences in candidate skills, and differences in passing rates can reduce the number of candidates hired. Second, virtually all the research to date on concerns with remote assessment has been on cognitive ability tests (e.g., Tippins et al. 2006; 2009), but there has been little on other common assessments that may be influenced by remote administration, such as case exercises and interviews that can be influenced by cheating, distractions, or potential loss of in-person information allowed by the context. This is especially problematic as surveys of employment assessments show that skill simulations are as common as employment tests, and interviews are much more common (e.g., Grossman 2024). Third, for organizations operating in diverse labor markets, adverse impact is a continual concern (Ployhart and Holtz 2008). Whether protected subgroups by racioethnicity[1] or gender perform less well with remote administration is of paramount importance, especially on cognitively loaded assessments because they tend to show larger racioethnic differences. Fourth, one might assume that the influence of remote assessments on candidate reactions is positive largely because they are more convenient for candidates compared with onsite assessments, but there has been little empirical research confirming this assumption (McCarthy et al. 2017). Because we lack studies in operational settings, we are also uncertain about differences in time periods, hiring procedures, and jobs that may confound comparisons, which are all addressed in the current study. Until these are investigated, proctored remote assessments are a risk as they are relatively costly compared with unproctored remote assessments and may result in other negative outcomes in terms of assessment quality, passing rates, adverse impact, and fairness perceptions compared with onsite assessments.

As such, the purpose of the current study is to evaluate potential differences in applicant scores on remote versus onsite assessments, whether remote assessments influence racioethnic and gender subgroup score differences, and how candidates react to the at-home testing environment. In doing so, we contribute to the literature in three ways. First, we draw from information processing theorizing and cognitive load research (Arthur, Keiser, Hagen, and Traylor 2018b; Arthur, Keiser, and Doverspike 2018a) to examine whether there are differences in scores for cognitive ability tests, case exercises, and structured interviews when conducted remotely versus onsite. This will contribute to the ongoing discussion on remote hiring and directly speak to any assessment-specific differences.

Second, we assess whether there are racioethnic or gender differences that may offer remote testing as an equalizing opportunity (Campion and Campion 2025). We also evaluate how candidates respond to the novel test-taking environment. Both are important to candidates and organizations. Organizations continue to focus on ways to manage subgroup differences in assessments, especially those correlated with cognitive ability in diverse labor markets. Further, candidates' experiences are especially relevant because their reactions are associated with many important outcomes such as organizational attraction, acceptance rates, referrals of other candidates, and many other behaviors (McCarthy et al. 2017).

Finally, although not a theoretical contribution but instead an empirical contribution, we conduct our study in a high-fidelity context. The current study examines differences in scores between onsite and proctored remote assessments and associated passing rates, including racial and gender subgroup differences, in an operational selection setting in the same organization, with large samples, during about the same time period, with the exact same set of representative hiring procedures, including cognitive ability tests, case exercises (skill simulations), and structured interviews for hiring for a wide range of professional jobs.

2 | Theoretical Background: Past Research on Remote Tests

The potential advantages of remote assessments, which have usually been discussed in the context of unproctored assessments, for hiring in terms of speed, cost, and flexibility were recognized more than 20 years ago (e.g., Mooney 2002). These benefits are a win-win because they are valued by hiring organizations and the candidates. Recent commentators describe remote assessments as a megatrend that cannot be avoided and instead must be intentionally considered when designing and validating modern selection systems (Jones and Cunningham 2023). This scholarship even suggests that the ease of availability of remote assessments might increase the number of candidates (Auer et al. 2022), thus improving the human capital potentially available to an organization. Notably, there is also some evidence of improved psychometrics (e.g., more variance and higher internal reliability; Ployhart et al. 2003).

However, there are many possible drawbacks to unproctored remote testing. The summaries of these concerns are well described from early in the trend (e.g., Tippins 2009; Tippins et al. 2006), including cheating, authentication of test-taker identity, the security of test content, and technology difficulties, which threaten the validity of the test scores, the utility of the tests, and professional ethical issues for high-stakes decisions. For example, Steger et al. (2020) meta-analyzed 109 effect sizes from 65 studies with a total sample of more than 100,000, showing an average difference in cognitive ability scores of 0.20 standard deviations (0.27 in high-stakes settings) such that unproctored settings had higher scores than proctored settings, illustrating the large effect of likely cheating on cognitive test scores. Although this meta-analysis did not explicitly assess remote versus onsite assessments, it nevertheless supports the continued pursuit of understanding environment as an influence on assessment scores.

The concerns over unproctored internet testing have greatly increased in the last 2 years with the advent of large language models (LLMs) such as ChatGPT that make cheating on unproctored tests easier and faster. Research examining this topic is just now emerging that documents how influential LLMs are in creating answers in laboratory settings. For example, Hickman et al. (2024) found ChatGPT performed well on verbal ability tests but not necessarily on quantitative ability tests; Harwood et al. (2024) found slightly higher scores on situational judgment tests; Canagasuriam and Lukacik (2025) found considerably higher scores on asynchronous video interviews; and Phillips and Robie (2024) found superior performance on personality assessments. The latter findings are particularly noteworthy because non-cognitive tests have been considered safer to use in unproctored settings because candidates can fake their answers whether unproctored or proctored. The consequence of LLMs is that the effect sizes for scores in unproctored conditions found in Steger et al.'s (2020) meta-analysis may increase because that study was before the popularity of LLMs. The widespread availability of AI tools may render any unproctored assessments ill-advised for high-stakes decisions.

Test developers and vendors have invented a number of ways to reduce the concerns with unproctored remote testing, such as retesting selected candidates later in proctored conditions to verify scores, limiting response time to not allow candidates the opportunity to look up answers, having a large number of items to reduce item-sharing concerns, only using unproctored methods for test types less susceptible to increased cheating given they can be faked whether in proctored or unproctored settings (e.g., personality or biodata, but see Phillips and Robie 2024), and many other partial solutions (e.g., Pearlman 2009; Tippins 2009; Tippins et al. 2006). However, the availability of LLMs may make many solutions ineffective.

Therefore, proctoring remote assessments may be a preferred solution since it solves most of these problems. Many approaches to remote proctoring have taken advantage of technology. For example, computer cameras allow observing the candidate during assessment as well as verifying identity, lockdown browsers can prevent candidates from looking up answers, and AI tools can monitor for any abnormalities in candidate behavior. In a recent review in education where proctored testing has been used most, Nigam et al. (2021) distinguish between live, recorded, and automated proctoring of remote tests. They identified and reviewed existing software and provided comparisons of the advantages and disadvantages of each. Research suggests that proctoring reduces cheating in educational settings (Alessio et al. 2017) and in simulated organizational settings with Mechanical Turk (MTurk) candidates (Karim et al. 2014). The latter study found negative reactions to proctoring, but that may have been due to priming effects because the experimental instructions explicitly emphasized the use of proctoring to reduce cheating. However, remote testing could potentially compensate for some cheating by generating an increased number of candidates due to easier access and then increasing the cutoff scores to pass the assessment. One simulation suggests if the increase in the applicant pool is high enough, performance outcomes improve despite increased cheating (Landers and Sackett 2012). Nevertheless, the potential for cheating with unproctored online assessments is still a great concern; thus, our focus on proctored assessments in the current study.

While proctored remote assessments may strike an ideal balance for many organizations between the pros and cons of unproctored remote assessments and onsite assessments, further work is needed to determine whether proctored assessments reasonably provide the positive outcomes organizations seek. Specifically, in this study, we test hypotheses that align with the aforementioned concerns: (1) whether the scores and passing rates will differ, (2) whether the outcomes will differ based on the type of assessments (e.g., tests, exercises, and interviews) compared with their onsite counterparts; (3) whether racioethnic subgroups differ by assessment environment; and (4) whether applicant fairness reactions differ by assessment environment.

3 | Hypothesis Development

3.1 | Differences by Types of Assessments

Whether scores on assessments differ is a major question for organizations that may consider implementing remotely proctored tests, especially since differences in subsequent passing rates (the number of candidates who meet an organization's cutoff for the test) can upset the number of candidates available for hire and the organization's ability to fulfill its human capital needs. This is especially the case for tests that are intended to capture cognitive ability and are administered unproctored, where candidates might look up answers or recruit assistance, as opposed to non-cognitive assessments (e.g., personality) where unproctored test scores tend to be equivalent (Steger et al. 2020), at least before LLMs. Remote proctoring should effectively reduce widespread cheating of this type, yet there likely still may be differences in scores on assessments completed remotely versus onsite.

We theorize that differences in scores may be due to the fact that the candidate must manage their finite cognitive resources differently by test and testing environment. Research on tests and assessments shows that types of assessments hold different amounts of "cognitive load"—or cognitive burden (Sweller et al. 1998). Cognitive ability tests are typically designed using multiple-choice items, which can be more cognitively demanding for candidates (Arthur et al. 2002). Meanwhile, constructed responses—or those where candidates generate their own response instead of selecting from pre-existing options (e.g., essays; Bennett 1991)—are shown to be less cognitively loaded (Edwards and Arthur 2007). This appears contrary to the educational testing literature that suggests test questions measuring recognition, such as multiple choice, require a lower level of understanding than recollection, such as constructed responses (e.g., Bloom's Taxonomy; Krathwohl 2002). However, with multiple-choice responses, candidates are expected to identify the single correct answer, whereas constructed responses allow for many correct answers and enable more creative interpretation where answers may not be viewed as simply "right" or "wrong." Moreover, test takers likely perceive opportunities to earn partial credit through demonstrating at least some knowledge. These differences could also influence the attitudes and perceptions of the test taker such that they may feel more test

anxiety when facing multiple-choice items where a "right" and a "wrong" exist than constructed response items where candidates have greater perceived latitude (Arthur et al. 2002; Edwards and Arthur 2007).

In addition to the cognitive loading of the method, research also shows that the technological design of the assessment influences performance (e.g., Landers and Marin 2021). Test takers have finite cognitive resources to spend on assessments, and the environment in which they complete the test can create an additional burden or not, as we discuss below.

3.1.1 | Cognitive Ability Tests

We expect that cognitive ability test performance is especially susceptible to environmental effects that strain information processing for several theoretical reasons. First, when completing a cognitive ability test remotely, candidates likely face greater distractions. Remote testing allows for what Arthur et al. (2018a) refer to as "high permissibility" contexts where candidates have volition in determining the location in which they complete the assessment. Remote candidates are encouraged to complete assessments in a quiet location without distractions, as is a best practice for administering hiring assessments (Society for Industrial and Organizational Psychology 2019, 37). Nevertheless, from the perspective of the organization, at-home testing environments are less controlled because they are unknown compared to onsite testing environments (e.g., presence of family members, house and street noise, pets and children, lack of quiet private location). Therefore, distractions due to remote testing can inhibit concentration and thus test performance. Second, while managing the cognitive load of the test, which is greater in cognitive ability tests than in other types of assessments, and managing potential distractions at home, candidates must also spend additional cognitive resources monitoring and managing their own behavior to not evoke accusations of cheating from the proctor who is observing them virtually (e.g., Stanton 2000).

In their initial empirical test of their information processing-based theorizing, Arthur et al.'s (2018b) found that remote testing may increase information processing requirements, thus reducing scores. Although they only compared scores on mobile devices versus desktops and not remote versus in-person, they found that participants scored lower on a test of general mental ability and that participants had to expend more cognitive effort on the more limiting device, which was the mobile device in their study. We argue that the concern with increased mental processing requirements will apply to test-takers in potentially distracting home environments versus onsite testing centers. As such, scores on remote cognitive ability tests should be lower than onsite administration due to the greater likelihood of distractions in the remote environment, the persistent demand for cognitive resources by the test method, and the fact that cognitive ability tests have a high cognitive load, creating a storm of cognitive depletion. Therefore, we propose:

Hypothesis 1a. *Cognitive ability test scores will be lower when administered in proctored remote conditions than when administered onsite.*

3.1.2 | Case Exercises

In contrast to cognitive ability tests, constructed responses, such as case exercises, require less cognitive load for candidates as they are allowed the flexibility to explain what they know in their own words. In the current context, the cases present hypothetical scenarios applicants would face on the job and prompts them to generate their own solution and course of action. As noted, constructed responses do not imply a "right" or "wrong" answer, but instead enable what Bennett (1991) calls "authentic assessments," which are intended to replicate the challenges and standards of performance that typically face members of a professional discipline (4). Moreover, the computer set-up is also more familiar (e.g., desk, keyboard, monitor, mouse, chair, lighting), and the candidate has extensive experience using their home computer. Nevertheless, for the same reasons as cognitive ability tests, we argue that a remote environment may still be more cognitive demanding given the potential for distraction and therefore yield lower scores when administered remotely versus onsite. We expect the cognitive load associated with the context to be the operative mechanism. Thus, we hypothesize:

Hypothesis 1b. *Case exercises scores will be lower when administered in proctored remote conditions than when administered onsite.*

3.1.3 | Structured Interviews

Because structured interviews also fall under the umbrella of constructed responses, we expect the same logic to apply as the case exercises, such that interviews have a lower cognitive load and virtual, video-based communication is a normal activity for nearly all contemporary professional workers, so completing such tasks in a home environment is less cognitively demanding than onsite. However, unlike case exercises, there appears to be a greater amount of published scholarship examining the differences in an interview environment. Empirical research comparing virtual versus in-person interview performance has yielded largely mixed findings, but these results require contextualization provided that many cannot be fairly compared to the present study. For example, in Blacksmith et al.'s (2016) meta-analysis of 12 independent samples testing the difference between virtual and in-person interviews, only one of their samples took place in an actual application context and included comparisons of interview ratings between videoconference and in-person interviews. The remaining samples were lab studies with university students (e.g., Straus et al. 2001), or if they did occur in the field, they compared in-person to telephone interviews (e.g., Silvester and Anderson 2003) or computer-mediated interviews that did not afford two-way communication (e.g., Thompson et al. 2007), or they compared applicant reactions and not interviewer ratings (e.g., Chapman et al. 2003). In the lab studies, the researchers' goal was to isolate the effects of remote media on performance, with the expected result that the media inhibited the interpersonal cues important in interviews, thus reducing scores. Further, some studies used student subjects as both interviewees and interviewers participating in mock interviews, and others conducted in the field with employed interviewers were not making actual hiring decisions. Real-world contexts

and consequences are influenced by many other factors that may be much more important than coping with the effects of technology. In actual selection contexts, with real hiring needs and especially in large organizations with public scrutiny of their hiring, other factors may change the expected results. The only study that occurred in a comparable context was conducted by Chapman and Rowe (2001) who found that ratings were higher in the virtual interviews than in-person interviews; and these results were not related to the degree of interview structure.

With few exceptions, published studies since this meta-analysis are similarly not fairly comparable. For example, in two studies on university students in simulated selection settings, Basch et al. (2021) and Melchers et al. (2021) found that interview performance ratings were lower in video-based interviews. These researchers attributed their findings to differences among medium to perceived social presence, perceived eye contact, and impression management. However, Langer et al. (2025) found small, nonsignificant mean differences between interview scores of in-person interviews versus a videoconference interview. Taking the modest existing empirical evidence together, we find that in comparable contexts, remote interviews are either rated higher or not meaningfully different from in-person interviews.

We believe at least two mechanisms help explain why we propose that remote proctored interview scores will be higher than onsite interview scores. The first is that video-based communication has become a standard practice in contemporary organizations, so applicants may have less test anxiety in a remote context. Further, completing such assessments in the comfort of one's home will likely further reduce any anxiety about the assessment, unlike in-person interviews that may feel less hospitable than one's home. Drawing from research on cognitive load and information processing theories, interviews may have a relatively lower cognitive load, and completing the assessment from home may require individuals to spend less time monitoring their environment because it is one with which they are already familiar; therefore, improving their performance on the remote interview.

An alternative—or perhaps co-occurring—mechanism is from the perspective of the raters. Some research suggests that humans may be likely to offer others the benefit of the doubt when the medium of communication is considered less rich (e.g., Short et al. 1976) and that they may attribute weaknesses in the interviewee's performance to the context rather than the interviewee's skills (e.g., Webster 1993). Moreover, applicants may perceive a disadvantage when completing remote assessments (Webster 1997) and, Interviewers may be sympathetic to these feelings or feel that they themselves would be disadvantaged if they were to be interviewed by videoconference (Chapman and Rowe 2001, 282). While these examples may appear to draw from an age when this technology was in its infancy and therefore technological effects would have been more impactful, the context of the present study mimics the unpredictability of these early experiences. The current study took place during the COVID-19 pandemic, which created global uncertainty that influenced the day-to-day experiences of hiring managers. These managers had to suddenly develop, implement, and monitor an online-only hiring system. It has long been recognized that hiring needs can greatly influence interviewer ratings (e.g., Carlson 1967), and COVID-19 created labor shortages across most industries and developed countries (Causa et al. 2022). Given this uncertainty, interviewers may have been sympathetic to the difficulties candidates experienced gave them the "benefit of the doubt" when making ratings.

Summarizing the empirical and theoretical evidence, we anticipate that remote structured interview scores will be higher than structured interviews administered onsite due to the interviewee's greater comfort in being interviewed from the comfort of their home, thus draining fewer cognitive resources than being interviewed onsite. Moreover, interviewers may be less stringent in their ratings of virtual interviewees for the reasons noted. As such, we hypothesize:

Hypothesis 1c. *Structured interview scores will be higher when administered remotely compared to onsite administration.*

3.2 | Differences by Racioethnic and Gender Subgroups

There are two ways we can assess the role of racioethnicity and gender in assessment score differences by testing environment. First, we can seek to understand whether there are within-group differences. This asks whether we would expect racioethnic minorities who complete assessments remotely to perform better or worse than their racioethnic counterparts who complete assessments onsite. Second, we can seek to understand whether there are any between-group differences. For large employers assessing a great number of candidates in diverse labor markets, adverse impact (i.e., differences in passing rates between subgroups) is an enduring legal concern, so changes in subgroup score differences by testing environment are of paramount importance. Any differences in subgroup by testing environment necessitate that changes occur within the subgroup. Put differently, for between-group differences to shrink between onsite and remote means that the racioethnic minority group scores would need to increase more or decrease less than nonminority group scores. In the test and assessment literature, there are no theories that would suggest that a non-minority group would perform worse than a racioethnic minority group in a different setting. However, research on stereotype threat suggests that individuals can be primed to perform in accordance with the stereotypes attached to their demographic features (Steele 1997). Thus, knowing the common racioethnic differences in test performance may hurt the performance of racioethnic minority candidates. In line with Hypotheses 1a and 1b, racioethnic minorities would face distractions and other cognitively depleting environmental stimuli just as racioethnic nonminorities would, yet the home environment would not trigger stereotype threat perceptions during the tests to the same extent as an onsite assessment center. This is because remote testing from home is a less priming environment than an office building where onsite testing would normally occur since the office location signals that the purpose is to be assessed. As such, racioethnic minority candidates will not feel as much stereotype threat from remote administration, thus reducing racioethnic differences in test

performance. Therefore, we propose that racioethnic minorities who take the assessments remotely will score better than those who take them onsite due to less stereotype threat from being remote (Hypothesis 2) and, because stereotype threat only applies to racioethnic minorities, these differences in scores will be greater than for non-minorities, thus reducing between-group relationships (Hypothesis 3).

Hypothesis 2. *Racioethnic minorities who complete the assessments remotely will score higher than racioethnic minorities who complete the assessments onsite for (a) cognitive ability tests, (b) case exercises, and (c) structured interviewers.*

Hypothesis 3. *Assessment score differences between racioethnic minorities and racioethnic non-minorities will be smaller for proctored remote assessments than onsite assessments for (a) cognitive ability tests, (b) case exercises, and (c) structured interviewers.*

Because these assessments did not include a mathematical component, we did not have reason to suggest gender differences (e.g., women scoring lower) by cognitive ability test or case exercise (Else-Quest et al. 2010). As such, we do not hypothesize within-gender (women remote vs. women onsite) or between-gender (women vs. men) differences, but instead pose them as a research question:

> Research Question: Are there (1) within-group or (2) between-group gender differences in scores for cognitive ability tests, case exercises, and interviews comparing remote and onsite administration?

3.3 | Differences in Applicant Fairness Reactions

Research suggests that candidate reactions to elements of the selection system, including assessments, can have a meaningful influence on crucial outcomes. For example, in their review of the candidate reaction literature, McCarthy et al. (2017) found candidate reactions to assessments were related to organizational attractiveness, a range of candidate intentions (e.g., to pursue the job, to accept a job if offered, to recommend other candidates), as well as test anxiety, test motivation, test performance, actual offer acceptance, and subsequent job performance. As such, how they respond to novel selection testing environments is essential for organizations to understand as they develop and implement these changes. Fortunately, reviews of the research have shown that candidates are generally favorable toward new technologies, as are organizations, provided there are advantages for both parties (Woods et al. 2020).

Extending this to remote assessments and incorporating the theoretical drivers from Hypothesis 1c, we argue that candidate reactions to remote assessments should generally be positive because they can complete these assessments from the comfort of their own home and computer, and do not need to take the time to travel to an onsite location that may feel sterile and more threatening. Few studies have examined this directly, and even fewer have occurred in comparable contexts. For example, Karim et al. (2014) compared the reactions of participants from MTurk responding to a cognitive ability test and found that those in the remote proctored condition had more negative test taker reactions and less cheating than those in the remote unproctored condition. However, in a study comparing remotely proctored and onsite testing for credentialing exams, Hurtz and Weiner (2022) found minimal differences in applicant reactions to either context. Remote respondents reported slightly lower satisfaction in testing software ease and onscreen instructions, were generally equivalent in proctor friendliness and helpfulness, and reported only marginally higher satisfaction with the noise level, compared to onsite respondents. Given the comfort afforded by a candidate's home environment, we expect applicants will report more positive reactions for the cognitive ability test and case exercises. Thus, we propose:

Hypothesis 4a. *Candidates will have more positive reactions to remote proctored assessments than onsite assessments for cognitive ability tests and case exercises.*

Counter to our theorizing for Hypothesis 4a, we expect candidates to have a more negative reaction to remote than onsite interviews. We propose this for at least two reasons. Primarily, though most workers have become accustomed to communicating virtually and the reduced social information available in a digital environment versus an in-person environment, remote interviewing may be viewed as less fair than onsite interviewing due largely to the high-stakes nature of employment testing (Gilliland 1993). Second, there is extensive evidence that candidates prefer in-person interviews to recorded video interviews (e.g., Basch and Melchers 2019; Chapman et al. 2003; Langer et al. 2017; Nørskov et al. 2020; Sears et al. 2013), including live video-conference interviews (Basch et al. 2021). Admittedly, recorded interviews are not comparable to the present context because these interviews were remote and live. Nevertheless, we expect that candidates would prefer to be in person to use their interpersonal skills and impression management techniques most effectively. Therefore, we hypothesize:

Hypothesis 4b. *Candidates will have more negative reactions to remote proctored assessments than onsite assessments for structured interviews.*

4 | Method

4.1 | Organizational Setting, Timeframe, and Sample

The data were collected in a single government organization in the United States between September 2019 and July 2022. This timeframe provided approximately the same number of assessments for analysis that were remote ($N = 902$) and onsite ($N = 891$). The remote assessments began in July 2020, as a response to restrictions on in-person interactions due to COVID-19, and assessments continued as remote-only until March 2021 when COVID-19 restrictions began to ease. From April to December 2021, assessments returned onsite, but then went back to remote-only assessments due to their positive

benefits perceived by the organization (e.g., saved cost of travel for candidates and increased flexibility for assessors) from January to August 2022, which was the end of data collection (except for candidate reactions, which were monitored for another year). One advantage of this study design is that there were no self-selection effects as candidates were not able to choose between remote and onsite assessments. This organization was considering whether to continue remote assessments permanently, which motivated the current study.

Data were collected across 19 professional jobs, meaning these jobs have specific skill requirements such as degrees and are salaried rather than paid hourly. Examples included jobs in information technology, finance, human resources, engineering, medicine, administration, and other more specialized jobs unique to the organization. The organization requested that the official job titles not be used in order to maintain anonymity. Candidates were hired by a centralized department in the organization, not by the managers to whom they would report, because they were assigned to a range of jobs throughout their typically long careers.

The database should be able to provide fairly accurate information about the differences between remote and onsite administration because the samples are large for both conditions, they were used in approximately the same time period, and both were used in the assessment of candidates applying to a wide range of jobs.

4.2 | Assessments

There were three assessments tailored to each job. First, there was a cognitive ability test consisting of job knowledge and/or a verbal skills test, depending on the job requirements, and a situational judgment test measuring the soft skills required by each job (which also tend to have some cognitive loading; McDaniel et al. 2001). Each section ranged in length from 30 to 50 items, with internal consistency reliabilities (Cronbach's alpha) generally exceeding 0.70. Candidates had 1 h to complete all three tests, and they were combined equally into a total score. The score was converted to the 7-point scale used for the other assessments, with equating based on using the same mean and standard deviation.

Second, there was a case exercise presenting a task for a new employee when entering the job, which was to develop a plan to address several existing problems described in a one-page description of the situation. The candidate wrote a 1–2 page memo to the supervisor describing the plan in 45 min. The memo was scored by two independent assessors on 7-point detailed anchored rating scales, which measured seven dimensions such as planning, problem solving, and writing skills, with both internal consistency reliabilities (Cronbach's alphas) and interrater reliabilities of the mean of the two assessors exceeding 0.70.[2]

Finally, there was a highly standardized three-part structured interview. The first part measured work experience and motivation for the job with three questions, the second part measured situational judgment with seven questions, and the third part measured past behavior with six questions. The situational and past behavior questions measured nine competencies, such as interpersonal skills, critical thinking, teamwork, and oral communication. The interviews lasted 1 h and were administered by a panel of two assessors who independently scored applicants on 7-point detailed behaviorally anchored rating scales tailored to each question. Both internal consistency reliabilities (Cronbach's alphas) and interrater reliabilities of the mean of the two assessors exceeded 0.70.

The assessments were developed based on job analyses and were content validated. The cognitive ability tests were based on blueprints (test specifications) deriving from the knowledge and skills important to each job based on job analyses, and the cases and interviews were based on about a dozen dimensions common to all jobs in the organization (e.g., teamwork, leadership, judgment, analysis, resourcefulness, writing skills). The assessors included one full-time assessor from Human Resources and one subject matter expert (SME) from the job, both of whom were highly trained and followed detailed scripts. The scores on the three tests were averaged equally to make a final score. Candidates scoring above a 5.25 passed and were rank ordered on a hiring list. Assessment scores determined hiring decisions directly. There were no exceptions or adjustments for labor shortages, but there were the typical disability accommodations (e.g., additional testing time, accessible information technology systems).

4.3 | Candidate Reactions Survey

All candidates were sent an anonymous survey (to ensure candidness) 1 week after their assessments to evaluate their reactions; thus, analyses will only be possible at a group level and not linked to individual test scores. We were able to compare candidate reactions in a time-series design, including before, during, and after the implementation of remote assessments. Moreover, the survey collected information on whether the candidate passed, the job that the candidate applied for, and the candidate's demographic statuses, which we used as controls.

We asked five questions about their perceived fairness of each of the assessments (based on Bauer et al. 2001): The exercise was related to the job to the best of my knowledge, I had an opportunity to show my skills through this exercise, The administration of the exercise was clear and transparent (e.g., understandable instructions, minimal distractions, few computer problems, etc.), I was treated with respect during this exercise, and Overall, I think the exercise was fair. All items were rated on a 5-point scale (strongly disagree—strongly agree) from which we formed a composite for each assessment that had internal consistency reliabilities (Cronbach's alphas) of 0.84 to 0.90. Candidates were also asked to report their racioethnicity, gender, the job they applied for, and whether or not they passed. We used the data starting 1 year before the study period to 1 year after the study period to have a sufficient time window and ensure a large enough sample to obtain stable estimates of any trends (from September 2018 to August 2023; $N_{remote} = 1{,}069$, and $N_{onsite} = 1{,}153$). The response rate was 67%.

Two questions were added to the survey to evaluate reactions to the remote testing: The remote administration of the assessments was easy to navigate and worked without difficulties and The remote administration of the assessments allowed me to participate more easily than in person. Both items were rated on the same 5-point agreement scale as described above. These were implemented in July 2023, which resulted in data from 74 candidates before the end of the data collection period.

4.4 | Procedures

All assessments were given to all candidates in the same order, with the cognitive ability test first, followed by the case exercise, and finally the interview. The organization in the current study used live remote proctoring by an administrative assistant via remote meeting software (e.g., Webex or Zoom) with highly detailed instructions and scripts to follow for monitoring to ensure standardization. Candidates were only allowed a laptop or desktop computer. This was a high-stakes setting for highly valued professional jobs where candidates were instructed and monitored not to have distractions, and candidates' internet connection was tested in advance. An individual proctor was assigned to each candidate to observe and monitor the candidate throughout the test and case assessments. The interviewers were present during the structured panel interviews so were naturally proctored. The cameras and the sound had to be on for candidates and proctors. Proctors were instructed to watch the candidate at every moment and to ask or remind the candidate about any unusual behavior, such as looking anywhere but on the main screen, using any other device, talking to anyone, and the like. They also monitored the candidate's screen via screen sharing so they would see if the candidate opened another browser or document. Candidates were instructed to only use one screen, and they would be observed looking at another screen if they did not follow instructions. The exam and case exercise did not allow the candidate to go to another browser or document because those options were removed by the software. They would have to exit the assessment, plus their screen was being monitored. Candidates were told that any sign of noncompliance would be grounds for disqualification. To verify identity, candidates had to show their driver's licenses to the proctor to be verified, and they had to show the proctor the room they were completing the assessment in before beginning.

Candidates were prescreened based on whether they met minimum job requirements for degrees and work experience, which were dichotomously scored by Human Resources staff. They were also prescreened based on the overall quality of their applications according to six skills required across jobs (e.g., job knowledge, communication skills, management skills, interpersonal skill), which were scored on 100-point scales by a panel of two SMEs from the jobs. The applications consisted of highly detailed forms to complete, including about six accomplishment record questions aligned with the skills above. The SMEs who scored the applications were incumbents who were independent of the assessors and each other. The most qualified candidates among those meeting minimum qualifications were selected to move forward to the focal assessments discussed in this study based on the number of candidates and hiring needs for each job. The purpose of the panel was to screen out unqualified and underqualified candidates unlikely to pass the assessments. The selection ratio at the prescreening stage varied by job and over time. Differences between remote and onsite assessments and across the phases of this study will be examined below. Although this prescreening likely reduced variance in subsequent assessment scores, such prescreening is virtually always necessary in any hiring context because a great number of candidates do not possess the minimum requirements or are not competitive for the jobs. Assessing candidates who do not have these minimum qualifications would be a waste of time and resources. Thus, this represents the situation that best reflects the context when organizations convert to remote testing. Differences in prescreening passing rates will be explored in the supplemental analysis section to determine if they provide an alternative explanation of the findings.

5 | Results

We first review the descriptive statistics to understand the basic data. We then test the hypotheses. We include supplemental analyses at the end to explore other possible reasons for the findings.

TABLE 1 | Samples by job.

Job	Frequency	Percent	Proportion remote
1	111	6.1	0.36
2	55	3.0	0.45
3	84	4.6	0.38
4	63	3.5	0.54
5	20	1.1	0.10
6	210	11.6	0.32
7	184	10.2	0.51
8	126	7.0	0.38
9	29	1.6	0.55
10	109	6.0	0.67
11	321	17.7	0.52
12	50	2.7	0.92
13	72	4.0	0.78
14	18	1.0	1.00
15	317	17.5	0.52
16	13	0.7	0.69
17	8	0.4	1.00
18	1	0.1	1.00
19	21	1.2	0.38
Total	1812	100.0	0.50

Note: The samples are a few candidates larger than the number that completed all assessments, as reported in the tables below.

5.1 | Descriptive Statistics

Table 1 shows that the samples included a large number of assessment scores for most jobs. The table also shows the proportion of assessment scores that were remote for each job. Most jobs have a good mix of remote and onsite assessment scores, with most proportions ranging from a third to two-thirds. However, some jobs' assessments have been all or mostly remote, such as Jobs 12, 14, and 18. This observation, along with differences by job in passing rates, led to the decision to control for job in all hypothesis tests. Table 2 shows that the samples were highly diverse. More than a third of the candidates are women and about half are from racial minority subgroups, with the largest groups being about 20% Hispanic or Latino, 13% Black, and 11% Asian. Racioethnic representation was virtually identical in both the onsite and remote conditions, differing by a few percent with no consistent trends. Table 3 shows that the average scores across assessments are in the low-to-mid-5 s on the 7-point scale, range from near 1 to near 7, have standard deviations of about two-thirds of a point, and are moderately intercorrelated. These data suggest that the tests differentiated well among the candidates and did not have excessive skew or restriction of range.

TABLE 2 | Samples by gender and race.

	Total		Onsite		Remote	
	n	%	*n*	%	*n*	%
Gender						
Women	657	36.5	334	37.4	323	35.7
Men	985	54.8	473	52.9	512	56.6
Missing	157	8.7	87	9.7	70	7.7
Total	1799	100	894	100	905	100
Race						
American Indian or Alaskan Native	15	0.8	6	0.7	9	1
Asian	205	11.4	93	10.4	112	12.4
Black	236	13.1	114	12.8	122	13.5
Hispanic	358	19.9	191	21.4	167	18.5
Native Hawaiian or other Pacific Islander	6	0.3	4	0.4	2	0.2
White	805	44.7	392	43.8	413	45.6
Missing	174	9.7	94	10.5	80	8.8
Total	1799	100	894	100	905	100

Note: Each candidate is counted in only one racial category. Hispanic candidates are counted in that category only, even if they indicated multiple races. White candidates include only those who indicated White and not multiple races. The samples are a few candidates larger than the number who completed all assessments as reported in the tables below.

5.2 | Tests of Hypotheses and Research Questions

We include analyses controlling for job because there were some differences in passing rates as well as hiring numbers across the phases of the data collection. We ran multiple regression models controlling for the differences between the 19 jobs by adding dummy variables that represented each job. We discuss the results with and without the controls because the covariation with job type means the differences may be due to jobs and not differences in remote administration, thus cannot be clearly interpreted. The univariate results are mainly for descriptive purposes, but we base our conclusions on the analyses with the controls.

Hypothesis 1a predicted that remote proctored administration of cognitive ability tests will have lower scores than onsite administration. Table 4 shows that remote tests had slightly lower scores ($d = -0.14$, $p < 0.01$), but were not significantly different when controlling for job. Hypothesis 1b predicted that remote proctored administration of case exercises will have lower scores than onsite administration. Table 4 shows there were no significant differences. Hypothesis 1c predicted that remote proctored administration of structured interviews will have higher scores than onsite administration. Table 4 shows both the mean differences ($d = 0.28$, $p < 0.001$) and standardized coefficient controlling for job differences ($\beta = 0.10$, $p < 0.001$) were significant. The total score also had

TABLE 3 | Descriptive statistics and intercorrelations on the tests.

						Intercorrelations		
Score	Mean	*N*	Standard deviation	Minimum	Maximum	Cognitive ability test	Case exercise	Structured interview
Cognitive ability test	5.35	1788	0.69	2.78	7.00			
Case exercise	5.14	1792	0.84	1.33	7.00	0.29*		
Structured interview	5.46	1791	0.57	1.81	6.83	0.18*	0.50*	
Final score	5.31	1793	0.54	1.39	6.57	0.66*	0.84*	0.71*

*$p < 0.001$ (2-tailed).

TABLE 4 | Remote versus onsite differences in test scores.

	Remote		On-site				
Score	**Mean**	***N***	**Mean**	***N***	**Difference**	***d***	***β***
Cognitive ability test	5.30	898	5.40	890	−0.10	−0.14**	0.00
Case exercise	5.17	902	5.12	890	0.04	0.06	0.01
Structured interview	5.54	901	5.38	890	0.16	0.28***	0.10***
Final score	5.33	902	5.29	891	0.04	0.07	0.05*
Passing	0.67	902	0.62	891	0.05		

Note: d = standardized mean difference. β = standardized regression coefficient controlling for job differences. *$p < 0.05$ (1-tailed), **$p < 0.01$ (1-tailed), and ***$p < 0.001$ (1-tailed).

TABLE 5 | Remote versus onsite differences by race and gender on the test scores.

		Cognitive ability test		Case exercise		Structured interview		Total score	
Race or gender	***N***	***d***	***β***	***d***	***β***	***d***	***β***	***d***	***β***
Asian	205	−0.32*	−0.03	0.19	0.14*	0.28*	0.17*	0.06	0.12*
Black	236	−0.19	−0.02	0.01	0.00	0.29*	0.18**	0.02	0.05
Hispanic	354	0.13	0.12**	0.21*	0.07	0.41***	0.16**	0.33***	0.16**
White	805	−0.10	0.05	0.03	0.02	0.23***	0.10**	0.06	0.07*
Women	657	0.01	0.09**	0.10	0.03	0.26***	0.11**	0.16*	0.10**
Men	985	−0.20***	−0.02	0.06	0.03	0.29***	0.12***	0.05	0.05
Total sample	1799	−0.14**	0.00	0.06	0.01	0.27***	0.10*	0.07	0.05*

Note: d = standardized mean difference. β = standardized regression coefficient controlling for job differences. The total sample includes those with missing race or gender. Positive values indicate higher scores on the remote assessments. *$p < 0.05$, (1-tailed), **$p < 0.01$ (1-tailed), and ***$p < 0.001$ (1-tailed).

a significant beta ($\beta = 0.05$, $p < 0.05$). Thus, only Hypothesis 1c was supported, indicating structured interview scores were higher in the remote condition.

Hypothesis 2a predicted that racioethnic minorities will have higher scores on the remotely administered cognitive ability test than on the onsite test. Table 5 shows that within group and controlling for job differences, only Hispanic or Latino candidates scored higher on the remote cognitive ability test compared to their onsite counterparts ($\beta = 0.12$, $p < 0.01$). However, when not controlling for job differences, Asian candidates who completed the cognitive ability test remotely fared worse than those who completed it onsite ($d = -0.31$, $p < 0.05$). As such, Hypothesis 2a is supported for only Hispanic or Latino candidates.

Hypothesis 2b predicted that racioethnic minorities will have higher scores on the remotely administered case exercise than the onsite test. Table 5 shows that within-group and controlling for job differences on Asian candidates scored higher on remotely administered case exercises compared to their onsite counterparts ($\beta = 0.14$, $p < 0.05$). As such, Hypothesis 2b is supported only for Asian candidates.

Hypothesis 2c predicted that racioethnic minorities will have higher scores on the remotely administered structured interview than those who completed it onsite. All racioethnic minority groups performed better on the remote structured interview ($\beta_{\text{Asian}} = 0.17$, $p < 0.05$, $\beta_{\text{Black}} = 0.18$, $p < 0.01$, $\beta_{\text{Hispanic/Latino}} = 0.16$, $p < 0.01$). As such, Hypothesis 2c is fully supported. Although not hypothesized, we also found that White candidates performed better on the structured interview administered remotely ($\beta = 0.10$, $p < 0.001$) than onsite.

Hypothesis 3a predicted that cognitive ability test score differences between racioethnic minorities and racioethnic nonminorities would be smaller when administered remotely versus onsite. We ran an ANCOVA for each comparison to control for job differences and found that there was a significant difference between Asian and White candidates comparing proctored remote versus onsite cognitive ability test scores ($F(1,988) = 3.30$, $p < 0.05$, 1-tailed)[3], but a graph showed that it was directionally counter to our hypothesis such that the difference in test scores was larger in the remote setting than onsite. As such, Hypothesis 3a is not supported.

Hypothesis 3b predicted that case exercise score differences would be smaller. We found no differences; thus, Hypothesis 3b is not supported. Hypothesis 3c predicted that structured interview score differences would be smaller. We found no differences; thus, Hypothesis 3c is not supported.

Research question 1 asked whether there are within-group gender differences in the assessments for remote versus onsite administration. We found that women performed better on the

remote cognitive ability test ($\beta = 0.09$, $p < 0.01$)and on the remote structured interview ($\beta = 0.11$, $p < 0.01$), compared with their onsite counterparts. Men scored better on the remote structured interviews ($\beta = 0.12$, $p < 0.001$) than their onsite counterparts. However, when not controlling for job differences, men who completed the cognitive ability test remotely fared worse than those who completed it onsite ($d = -0.20$, $p < 0.001$). Research question 2 asked whether there were between-group gender differences. We found score differences between women and men for cognitive ability tests ($F(1,1620) = 4.92$, $p < 0.05$). The interaction for cognitive ability tests was ordinal such that while women performed less well than men onsite, they outperformed men remotely.

Hypothesis 4a predicted that candidates would have a more positive reaction to remotely administered assessments than onsite assessments for cognitive ability tests and case exercises, and Hypothesis 4b predicted that candidates would have more negative reactions to remotely administered structured interviews than onsite interviews. Table 6 shows that the mean ratings were virtually identical across time periods for each assessment, and none of the ANCOVAs across time periods controlling for the candidate's demographic statuses, the job that the candidate applied for, and whether the candidate passed were significant for any assessment. Thus, Hypotheses 4a and 4b were not supported.

5.3 | Supplemental Analysis

Supplemental analyses explored two alternative explanations of the findings. First, we examined other indicators of candidate quality across remote and onsite administration periods. Specifically, it was conceivable that candidates seeking employment during COVID-19 were more motivated to get a better job or were laid off due to COVID-19, given the greater difficulty and health risks. As noted in the Section 4, candidates were preselected based on meeting minimum qualifications, a review, and scoring of applications and accomplishment records by a panel of two subject matter experts. The most qualified candidates were selected to move forward to the assessments discussed in this study based on the number of candidates and hiring needs for each job. The selection ratio at the prescreening stage varied by job and over time, thus possibly providing insight as to the reasons for the results.

Table 7 shows the passing rates on the two components of the prescreen: meeting minimum qualifications and overall quality ratings. The average number of candidates per hiring wave dropped initially when going remote, continued to drop when going back onsite, but then began to recover somewhat when going permanently remote. There was an increase in meeting minimum qualifications when initially moving to remote, suggesting higher-quality candidates based on possessing the required degrees and work experience, which receded when administration went back to onsite, but then increased when going permanently remote, with a higher average for remote. To the contrary, there was a drop in passing rates based on candidate quality ratings when going remote initially, suggesting lower-quality candidates on dimensions related to the assessments, which recovered a bit when back onsite (sample-weighted

TABLE 6 | Remote versus onsite differences in candidate reactions.

Time period	*N*	Cognitive tests	Assessment exercises (cases)	Structured interviews	Total
Onsite testing (September 2018 to June 2020)	695	4.37	4.37	4.35	4.36
Remote testing (July 2020 to March 2021)	155	4.34	4.33	4.32	4.33
Onsite testing (April 2021 to December 2021)	374	4.35	4.53	4.34	4.35
Remote testing (January 2022 to August 2023)	998	4.38	4.38	4.37	4.36

Note: None of the ANCOVAs across time periods controlling for whether the candidate passed, the job that the candidate applied for, and the candidate's demographic statuses were significant ($p < 0.05$, 1-tailed) for any assessments.

TABLE 7 | Changes in minimum qualifications and application quality prescreen passing rates by stage.

	Average *n* by job		Average passing rate across jobs		Sample weighted passing rate	
Stage	Meeting minimum qualifications	Ratings of application quality	Meeting minimum qualifications	Ratings of application quality	Meeting minimum qualifications	Ratings of application quality
Pre remote	130.07	40.63	0.60	0.46	0.31	0.47
First remote	93.16	46.63	0.71	0.37	0.50	0.30
Back to onsite	63.84	27.11	0.58	0.32	0.42	0.33
Back to remote	75.10	40.08	0.46	0.54	0.49	0.50

passing rates), but then recovered after going permanently remote, with no average difference between remote and onsite. Thus, the two indicators are somewhat contradictory and do not suggest that candidates were of either lower or higher quality during remote administration.

In a second supplemental analysis, we rationally evaluated whether the higher passing rates for the structured interviews were due to the interviewers making up for hiring shortages or due to giving candidates the "benefit of the doubt" by making higher ratings because of the difficulties of doing the interviews remotely. We found that the policy of the organization was not to give interviewers this information so as to not bias their judgments, plus the organizational unit that assessed the candidates in this large government organization was separate from the unit that drew candidates from those passing to enter training based on hiring needs. Thus, the assessment unit was not aware of any hiring shortages if they existed. Thus, they would not be able to communicate this information to interviewers. Whether interviewers made assumptions about shortages or gave candidates the benefit-of-the-doubt is not known.

6 | Discussion

A critical issue for organizations is whether hiring candidates remotely will influence the candidates available for hire to meet staffing needs. The present study provided a strong test by comparing onsite assessments to remotely proctored assessments in an operational setting in a large organization for a wide range of professional jobs using three types of common selection procedures at about the same time period. The results indicated no differences for cognitive ability tests or case exercises, but higher scores for structured interviews, leading to higher total scores for all the assessments combined. Within-group comparisons show that all subgroups had higher interview scores remotely, and Hispanic or Latino and women candidates also had higher cognitive test scores remotely, compared to their racioethnic and gender counterparts onsite. Between-group racioethnic differences were not smaller in the remote condition, and Asian candidates appeared to perform worse on the cognitive ability test in the remote condition compared to White candidates. However, women appeared to perform better than men on the cognitive ability test in the remote condition. Finally, candidate fairness reactions did not differ for remote compared with onsite for any assessments or any subgroups.

6.1 | Theoretical Implications

One primary theoretical implication is that there were no consistent differences in cognitive ability test or case exercise scores despite potential reasons why remote proctored administration might either yield lower scores due to distractions in the remote setting, especially for cognitive tests (e.g., Arthur, Keiser, and Doverspike 2018a; Arthur, Keiser, Hagen, and Traylor 2018b; Society for Industrial and Organizational Psychology 2019), or higher scores due to greater comfort from using one's home computer for a test involving extensive typing such as case exercises (e.g., Choi et al. 2014; Hancock 2001). The implication is that these theoretical reasons for differences may not matter and therefore suggests that proctored test environments may evoke the same kind of experience as onsite proctored environments for test takers completing a cognitive ability test or case exercise.

Similarly important for theory, scores on remote structured interviews were higher ($d = 0.28$) despite extensive research showing lower scores for remote administration (e.g., Basch et al. 2021; Blacksmith et al. 2016; Melchers et al. 2021). Given the limitations of existing scholarship occurring in a lab rather than in operational settings and testing unproctored rather than proctored assessments, we theorized and found that proctored remote interviews in actual selection settings would yield higher scores than on-site interviews (e.g., Langer et al. 2025). A theoretical implication is that proctored remote interviews allow candidates to use interpersonal skills, the interviewers to detect those skills, and there can be a live exchange, such that the scores do not degrade. Likewise, both the candidate and the interviewers may be more motivated to overcome the limitations of remote testing if actual hiring is at stake. The theoretical reasons for differences in remote assessments in past research have focused almost exclusively on the influence on the candidates, but it may be just as relevant to consider potential rater effects.

For example, our finding of no differences in the cognitive ability tests and case exercises may be because they are highly objective in administration and scoring; thus, the setting does not matter. On the other hand, the structured interviews have a subjective component. Despite the highly structured interview process with standardized questions, rating scales, multiple independent interviews, and interviewer training, human judgment may have considered the change in the administration and awarded higher scores in remote conditions. This appears not to be due to differences in the quality of candidates, overall passing rates, or knowledge of hiring needs. Future research should explore directly whether interviewers compensate or give candidates higher ratings in remote administration to make up for the limitations of the media. As such, future research should further examine how the organizational context may play a role.

There were few within-group racioethnic differences despite theoretical differences to expect otherwise, such as reducing stereotype threat by completing assessments from home (Steele 1997), or the influence of compensatory equalization by raters who may offer the benefit of the doubt to remote candidates. Again, an important theoretical implication may be that the test environment is not a boundary condition of stereotype threat. It may be that the proctored nature of the remote assessments triggers the same kind of stereotype threat that an onsite assessment environment would. Contrary to predictions, Hispanic or Latino and women candidates scored higher within the group compared to White and male candidates on the cognitive tests when administered remotely. The theoretical explanation for this latter finding, if any, is unclear.

6.2 | Practical Implications

There are several practical implications from this research. First, remote assessments are cost-effective and convenient;

thus, selection experts must figure out how to implement them without suffering distortion from cheating. A solution is to use proctoring. Although this adds additional expense, it is less expensive than onsite administration for the organization (e.g., office space) and candidate (e.g., travel time and costs). A key concern prior to this study was whether it would influence scores and passing rates. We found that going remote may not have a detrimental effect on overall scores. In fact, we found scores and passing rates may increase slightly. To illustrate the practical effects, the 5% increase in overall passing rate would have added about 17 more passing candidates to the 552 that passed if the 891 onsite assessments had been administered remotely in proctored conditions. Although seemingly a small gain, the financial value is large. Hires normally stay for 20-year careers. Using the average wages, including benefits and retirement, each hire is a 4-million-dollar investment. Other organizations may have different experiences depending on their hiring strategy. For example, in the current study, all candidates were given all three assessments, and the scores were combined equally. The practical implications might be different with other strategies, such as a multiple-hurdle process where a less expensive cognitive test might be used as the first stage. The slightly lower scores for remote administration could reduce the numbers moving to the latter stages a small amount, but that may depend on the candidates.

Second, there are many types of remote assessments. It is not just a concern for traditional employment tests like cognitive ability tests, but interviews and case exercises are also of interest. Exercises could be influenced by cheating, but interviews might be influenced by the dynamics of face-to-face versus teleconference media. The current study finds little difference in scores with remote proctoring versus onsite administration of case exercises and score improvements with interviews.

Third, subgroup differences are a concern in many contexts, such as the U.S. and many other countries, due to antidiscrimination employment laws. The current study suggests that proctored remote assessments do not worsen subgroup differences and subsequent adverse impacts, and it may even help representation. There was some evidence that the type of assessment might matter. Hispanic or Latino and women candidates performed higher on remote cognitive tests, and men candidates performed lower on cognitive tests, but the difference was not significant when job type was controlled. Racioethnic minority subgroups showed greater within-group score increases in the interview than racioethnic non-minorities. Moreover, although Table 2 showed that the relative diversity of the candidates did not differ over the course of this study, remote testing may improve the diversity of the candidate pool because it makes the assessments much more accessible (but see Auer et al. 2022).

Finally, candidate reactions are an extremely important outcome aside from scores because candidate reactions may influence acceptance rates, reapplication decisions, referrals to other candidates, discrimination perceptions, the reputation of the employer, and possible future customers. The current study shows that candidate fairness reactions to assessments are not affected by remote versus onsite administration, and candidates appear to recognize the benefits.

6.3 | Limitations and Future Research

One area for future research is to further examine the role of subjectivity in understanding remote differences in assessment scores, and particularly in the interview. Another area is to examine potential reasons for the differences in scores between proctored remote and onsite testing that were observed here but not explained, such as Hispanic or Latino and women candidates performing better on cognitive tests administered remotely. Future research might examine the influence of proctoring on other types of assessments where cheating is possible but has not yet risen to the attention of organizations and researchers, such as using LLMs to generate resumes, application information, answers to open-ended questions, and other text-based assessments in usually unproctored conditions. Another limitation is that the study did not evaluate the criterion-related validity of the scores, which is an especially important outcome for future research. Finally, as always, future research should replicate these findings in other contexts with different organizations, assessments, and jobs. As noted previously, the organizational context may play a role, such as the need for candidates or sympathy for the remote conditions. As another example, would only assessing some candidates remotely, such as if they would have to travel, or allowing the candidate the choice to take assessments live or remotely, influence the results?

Data Availability Statement

Research data are not shared.

Endnotes

[1] In line with APA guidelines (https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/racial-ethnic-minorities) to not confound race and ethnicity by referring to only "race," we adopt the term "racioethnicity" (also see Cox 2004; Nkomo and Hoobler 2014).

[2] For both the cases and the interviews, inter-rater reliability statistics were calculated using raters' mean scores (ICC2). Disagreements between raters were resolved by discussing until they reached consensus within one point of each other.

[3] We are not able to report the exact mean differences for proprietary reasons. Only significance tests are reported.

References

Alessio, H. M., N. Malay, K. Maurer, A. J. Bailer, and B. Rubin. 2017. "Examining the Effect of Proctoring on Online Test Scores." *Online Learning* 21, no. 1: 146–161.

Arthur, W., B. D. Edwards, and G. V. Barrett. 2002. "Multiple-Choice and Constructed Response Tests of Ability: Race-Based Subgroup Performance Differences on Alternative Paper-And-Pencil Test Formats." *Personnel Psychology* 55, no. 4: 985–1008. https://doi.org/10.1111/j.1744-6570.2002.tb00138.x.

Arthur, W., N. L. Keiser, and D. Doverspike. 2018a. "An Information-Processing-Based Conceptual Framework of the Effects of Unproctored Internet-Based Testing Devices on Scores on Employment-Related Assessments and Tests." *Human Performance* 31, no. 1: 1–32. https://doi.org/10.1080/08959285.2017.1403441.

Arthur, W., N. L. Keiser, E. Hagen, and Z. Traylor. 2018b. "Unproctored Internet-Based Device-Type Effects on Test Scores: The Role of Working

Memory." *Intelligence* 67: 67–75. https://doi.org/10.1016/j.intell.2018.02.001.

Auer, E. M., K. J. Cavanaugh, J. R. Petor, T. B. Kinney, and R. N. Landers. 2022. "The Effects of Unproctored Internet Testing on Applicant Pool Size and Diversity: Using Interrupted Time Series to Improve Causal Inference." *Technology, Mind, and Behavior* 3. https://doi.org/10.1037/tmb0000079.

Basch, J. M., and K. G. Melchers. 2019. "Fair and Flexible?! Explanations Can Improve Applicant Reactions Toward Asynchronous Video Interviews." *Personnel Assessment and Decisions* 5, no. 3: 1–11. https://doi.org/10.25035/pad.2019.03.002.

Basch, J. M., K. G. Melchers, A. Kurz, M. Krieger, and L. Miller. 2021. "It Takes More Than a Good Camera: Which Factors Contribute to Differences Between Face-To-Face Interviews and Videoconference Interviews Regarding Performance Ratings and Interviewee Perceptions?" *Journal of Business and Psychology* 36: 921–940. https://doi.org/10.1007/s10869-020-09714-3.

Bauer, T. N., D. M. Truxillo, R. J. Sanchez, J. M. Craig, P. Ferrara, and M. A. Campion. 2001. "Applicant Reactions to Selection: Development of the Selection Procedural Justice Scale (SPJS)." *Personnel Psychology* 54: 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x.

Bennett, R. E. 1991. "On the Meanings of Constructed Response." *ETS Research Report Series* 1991, no. 2: i–46.

Blacksmith, N., J. C. Willford, and T. S. Behrend. 2016. "Technology in the Employment Interview: A Meta-Analysis and Future Research Agenda." *Personnel Assessment and Decisions* 2, no. 1: 12–20. https://doi.org/10.25035/pad.2016.002.

Campion, E. D., and M. A. Campion. 2025. "Using Practice Employment Tests in Recruitment and Selection to Equaliz Preparation Opportunities." *Human Resource Management* 64, no. 3: 879–899.

Canagasuriam, D., and E. R. Lukacik. 2024. "ChatGPT, Can You Take My Job Interview? Examining Artificial Intelligence Cheating in the Asynchronous Video Interview." *International Journal of Selection and Assessment* 33, no. 1. https://doi.org/10.1111/ijsa.12491.

Carlson, R. E. 1967. "Selection Interview Decisions: The Effect of Interviewer Experience, Relative Quota Situation, and Applicant Sample on Interviewer Decisions." *Personnel Psychology* 20: 259–280. https://doi.org/10.1111/j.1744-6570.1967.tb01523.x.

Causa, O., M. Abendschein, N. Luu, E. Soldani, and C. Soriolo. 2022. "The Post-COVID-19 Rise in Labour Shortages." *Organization for Economic Cooperation and Development*. https://doi.org/10.1787/e60c2d1c-en.

Chapman, D. S., and P. M. Rowe. 2001. "The Impact of Videoconference Technology, Interview Structure, and Interviewer Gender on Interviewer Evaluations in the Employment Interview: A Field Experiment." *Journal of Occupational and Organizational Psychology* 74: 279–298. https://doi.org/10.1348/096317901167361.

Chapman, D. S., K. L. Uggerslev, and J. Webster. 2003. "Applicant Reactions to Face-To-Face and Technology-Mediated Interviews: A Field Investigation." *Journal of Applied Psychology* 88, no. 5: 944–953. https://doi.org/10.1037/0021-9010.88.5.944.

Choi, H. H., J. J. Merriënboer, and F. Paas. 2014. "Effects of the Physical Environment on Cognitive Load and Learning: Towards a New Model of Cognitive Load." *Educational Psychology Review* 26: 225–244.

Cox, T., Jr. 2004. "Problems with Research by Organizational Scholars on Issues of Race and Ethnicity." *Journal of Applied Behavioral Science* 40, no. 2: 124–145.

Edwards, B. D., and W. Arthur. 2007. "An Examination of Factors Contributing to a Reduction in Subgroup Differences on a Constructed-Response Paper-And-Pencil Test of Scholastic Achievement." *Journal of Applied Psychology* 92, no. 3: 794.

Else-Quest, N. M., J. S. Hyde, and M. C. Linn. 2010. "Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis." *Psychological Bulletin* 136, no. 1: 103.

Gilliland, S. W. 1993. "The Perceived Fairness of Selection Systems: An Organizational Justice Perspective." *Academy of Management Review* 18, no. 4: 694–734.

Grossman, K. 2024. "2023 Global Candidate Experience (CandE) Benchmark Research Report ERE Media." https://www.eremedia.com/reports/2023-global-candidate-experience-cande-benchmark-research-report.

Hancock, D. R. 2001. "Effects of Test Anxiety and Evaluative Threat on students' Achievement and Motivation." *Journal of Educational Research* 94: 284–290. https://doi.org/10.1080/00220670109598764.

Harwood, H., N. Roulin, and M. Z. Iqbal. 2024. "Anything You Can Do, I Can Do: Examining the Use of ChatGPT in Situational Judgement Tests for Professional Program Admission." *Journal of Vocational Behavior* 154: 104013. https://doi.org/10.1016/j.jvb.2024.104013.

Hickman, L., P. D. Dunlop, and J. L. Wolf. 2024. "The Performance of Large Language Models on Quantitative and Verbal Ability Tests: Initial Evidence and Implications for Unproctored High-Stakes Testing." *International Journal of Selection and Assessment* 32: 499–511. https://doi.org/10.1111/ijsa.12479.

Hurtz, G. M., and J. A. Weiner. 2022. "Comparability and Integrity of Online Remote vs. Onsite Proctored Credentialing Exams." *Journal of Applied Testing Technology*: 36–45.

Jones, J. W., and M. R. Cunningham. 2023. "Going Beyond a Validity Focus to Accommodate Megatrends in Selection System Design." *Industrial and Organizational Psychology* 16, no. 3: 336–340. https://doi.org/10.1017/iop.2023.28.

Karim, M. N., S. E. Kaminsky, and T. S. Behrend. 2014. "Cheating, Reactions, and Performance in Remotely Proctored Testing: An Exploratory Experimental Study." *Journal of Business and Psychology* 29: 555–572. https://doi.org/10.1007/s10869-014-9343-z.

Krathwohl, D. R. 2002. "A Revision Bloom's Taxonomy: An Overview." *Theory Into Practice* 41, no. 4: 212–218. https://www.jstor.org/stable/1477405.

Landers, R. N., and S. Marin. 2021. "Theory and Technology in Organizational Psychology: A Review of Technology Integration Paradigms and Their Effects on the Validity of Theory." *Annual Review of Organizational Psychology and Organizational Behavior* 8, no. 1: 235–258.

Landers, R. N., and P. R. Sackett. 2012. "Offsetting Performance Losses due to Cheating in Unproctored Internet-Based Testing by Increasing the Applicant Pool." *International Journal of Selection and Assessment* 20, no. 2: 220–228. https://doi.org/10.1111/j.1468-2389.2012.00594.x.

Langer, M., A. Demetriou, A. Arvanitidis, S. Vanderveken, and A. M. Hiemstra. 2025. "A Quasi-Experimental Investigation of Differences Between Face-To-Face and Videoconference Interviews in an Actual Selection Process." *Applied Psychology* 74, no. 1. https://doi.org/10.1111/apps.12558.

Langer, M., C. J. König, and K. Krause. 2017. "Examining Digital Interviews for Personnel Selection: Applicant Reactions and Interviewer Ratings." *International Journal of Selection and Assessment* 25, no. 4: 371–382. https://doi.org/10.1111/apps.12558.

Maurer, R. 2021. "With Virtual Interviews Here to Stay, Best Practices Are Needed." *Society of Human Resource Management*. https://www.shrm.org/topics-tools/news/talent-acquisition/virtual-interviews-to-stay-best-practices-needed.

McCarthy, J. M., T. N. Bauer, D. M. Truxillo, N. R. Anderson, A. C. Costa, and S. M. Ahmed. 2017. "Applicant Perspectives During Selection: A Review Addressing "So What?," "What's New?," and "Where to Next?""

Journal of Management 43, no. 6: 1693–1725. https://doi.org/10.1177/0149206316681846.

McDaniel, M. A., F. P. Morgeson, E. B. Finnegan, M. A. Campion, and E. P. Braverman. 2001. "Use of Situational Judgment Tests to Predict Job Performance: A Clarification of the Literature." *Journal of Applied Psychology* 86: 730–740. https://doi.org/10.1037/0021-9010.86.4.730.

Melchers, K. G., A. Petrig, J. M. Basch, and J. Sauer. 2021. "A Comparison of Conventional and Technology-Mediated Selection Interviews With Regard to interviewees' Performance, Perceptions, Strain, and Anxiety." *Frontiers in Psychology* 11: 603–632. https://doi.org/10.1016/j.hrmr.2020.100789.

Mooney, J. 2002. "Pre-Employment Testing on the Internet: Put Candidates a Click Away and Hire at Modem Speed." *Public Personnel Management* 31, no. 1: 41–52. https://doi.org/10.1177/009102600203100105.

Nigam, A., R. Pasricha, T. Singh, and P. Churi. 2021. "A Systematic Review on AI-Based Proctoring Systems: Past, Present and Future." *Education and Information Technologies* 26, no. 5: 6421–6445. https://doi.org/10.1007/s10639-021-10597-x.

Nkomo, S., and J. M. Hoobler. 2014. "A Historical Perspective on Diversity Ideologies in the United States: Reflections on Human Resource Management Research and Practice." *Human Resource Management Review* 24, no. 3: 245–257.

Nørskov, S., M. F. Damholdt, J. P. Ulhøi, M. B. Jensen, C. Ess, and J. Seibt. 2020. "Applicant Fairness Perceptions of a Robot-Mediated Job Interview: A Video Vignette-Based Experimental Survey." *Frontiers in Robotics and AI* 7: 58626. https://doi.org/10.3389/frobt.2020.586263.

Parket, K., J. Horowitiz, and R. Minkin. 2022. Covid-19 pandemic continues to reshape work in America Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2022/02/PSDT_2.16.22_covid_work_report_clean.pdf.

Pearlman, K. 2009. "Unproctored Internet Testing: Practical, Legal, and Ethical Concerns." *Industrial and Organizational Psychology* 2, no. 1: 14–19.

Phillips, J., and C. Robie. 2024. "Can a Computer Outfake a Human?" *Personality and Individual Differences* 217: 112434. https://doi.org/10.1016/j.paid.2023.112434.

Ployhart, R. E., and B. C. Holtz. 2008. "The Diversity–Validity Dilemma: Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact in Selection." *Personnel Psychology* 61, no. 1: 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x.

Ployhart, R. E., J. A. Weekley, B. C. Holtz, and C. Kemp. 2003. "Web-Based and Paper-And-Pencil Testing of Applicants in a Proctored Setting: Are Personality, Biodata, and Situational Judgment Tests Comparable?" *Personnel Psychology* 56, no. 3: 733–752. https://doi.org/10.1111/j.1744-6570.2003.tb00757.x.

Sears, G. J., H. Zhang, W. H. Wiesner, R. D. Hackett, and Y. Yuan. 2013. "A Comparative Assessment of Videoconference and Face-To-Face Employment Interviews." *Management Decision* 51, no. 8: 1733–1752. https://doi.org/10.1108/MD-09-2012-0642.

Short, J., E. Williams, and B. B. Christie. 1976. *The Social Psychology of Telecommunications*. Wiley.

Silvester, J., and N. Anderson. 2003. "Technology and Discourse: A Comparison of Face-To-Face and Telephone Employment Interviews." *International Journal of Selection and Assessment* 11, no. 2-3: 206–214.

Society for Industrial and Organizational Psychology. 2019. *Principles for the Validation and Use of Personnel Selection Procedures*. 5th ed. Society for Industrial and Organizational Psychology.

Stanton, J. M. 2000. "Reactions to Employee Performance Monitoring: Framework, Review, and Research Directions." *Human Performance* 13, no. 1: 85–113. https://doi.org/10.1207/S15327043HUP1301_4.

Steele, C. M. 1997. "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance." *American Psychologist* 52, no. 6: 613–629. https://doi.org/10.1037/0003-066X.52.6.613.

Steger, D., U. Schroeders, and T. Gnambs. 2020. "A Meta-Analysis of Test Scores in Proctored and Unproctored Ability Assessments." *European Journal of Psychological Assessment* 36, no. 1: 174–184. https://doi.org/10.1027/1015-5759/a000494.

Straus, S. G., J. A. Miles, and L. L. Levesque. 2001. "The Effects of Videoconference, Telephone, and Face-To-Face Media on Interviewer and Applicant Judgments in Employment Interviews." *Journal of Management* 27, no. 3: 363–381.

Sweller, J., J. J. G. Van Merriënboer, and F. Paas. 1998. "Cognitive Architecture and Instructional Design." *Educational Psychology Review* 10: 251–295.

Thompson, L. F., E. A. Surface, and T. J. Whelan. 2007. "Examinees' Reactions to Computer-Based Versus Telephonic Oral Proficiency Interviews." In *Paper presented at the 22nd Annual Conference of the Society of Industrial and Organizational Psychology.*

Tippins, N. T. 2009. "Internet Alternatives to Traditional Proctored Testing: Where Are We Now?" *Industrial and Organizational Psychology* 2, no. 1: 2–10. https://doi.org/10.1111/j.1754-9434.2008.01097.x.

Tippins, N. T. 2015. "Technology and Assessment in Selection." *Annual Review of Organizational Psychology and Organizational Behavior* 2, no. 1: 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317.

Tippins, N. T., J. Beaty, F. Drasgow, et al. 2006. "Unproctored Internet Testing in Employment Settings." *Personnel Psychology* 59, no. 1: 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x.

Webster, D. M. 1993. "Motivated Augmentation and Reduction of the Over Attribution Bias." *Journal of Personality and Social Psychology* 65: 261–271.

Webster, J. 1997. Selection Interviews Through Video Conferencing: Interviewees' Reactions. Paper Presented at the 1997 Academy of Management Meetings.

Woods, S. A., S. Ahmed, I. Nikolaou, A. C. Costa, and N. R. Anderson. 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, no. 1: 64–77. https://doi.org/10.1080/1359432X.2019.1681401.