International Journal of Selection and Assessment

WILEY

International Journal of SELECTION AND ASSESSMENT

RESEARCH ARTICLE OPEN ACCESS

Can Interviewees Fake Out AI? Comparing the Susceptibility and Mechanisms of Faking Across Self-Reports, Human Interview Ratings, and AI Interview Ratings

Louis Hickman[1] | Josh Liff[2] | Colin Willis[2] | Emily Kim[1]

[1]Virginia Tech, Blacksburg, Virginia, USA | [2]HireVue Inc., South Jordan, Utah, USA

Correspondence: Louis Hickman (louishickman@vt.edu)

Received: 24 February 2025 | **Revised:** 22 April 2025 | **Accepted:** 24 April 2025

ABSTRACT

Artificial intelligence (AI) is increasingly used to score employment interviews in the early stages of the hiring process, but AI algorithms may be particularly prone to interviewee faking. Our study compared the extent to which people can improve their scores on self-report scales, structured and less structured human interview ratings, and AI interview ratings. Further, we replicate and extend prior research by examining how interviewee abilities and impression management tactics influence score inflation across scoring methods. Participants ($N = 152$) completed simulated, asynchronous interviews in honest and applicant-like conditions in a within-subjects design. The AI algorithms in the study were trained to replicate question-level structured interview ratings. Participants' scores increased most on self-reports (overall Cohen's $d = 0.62$) and least on AI interview ratings (overall Cohen's $d = 0.14$), although AI score increases were similar to those observed for human interview ratings (overall Cohen's $d = 0.22$). On average, across conditions, AI interview ratings converged more strongly with structured human ratings based on behaviorally anchored rating scales than with less structured human ratings. Verbal ability only predicted score improvement on self-reports, while increased use of honest defensive impression management tactics predicted improvement in AI and less structured human interview scores. Ability to identify criteria did not predict score improvement. Overall, these AI interview scores behaved similarly to structured human ratings. We discuss future possibilities for investigating faking in AI interviews, given that interviewees may try to "game" the system when aware that they are being evaluated by AI.

Natural language processing (NLP) and machine learning (ML) are increasingly used as an artificial intelligence (AI) approach to scoring open-ended assessments, including employment interviews (Hickman et al. 2022; Liff et al. 2024). AI scored interviews are generally deployed in the early stages of the hiring process as an alternative to other inexpensive assessments, such as personality self-reports (Hickman et al. 2025), or as a scalable approach for supplementing traditional structured interviews conducted during the hiring process (Liff et al. 2024). One motivation for doing so is to reduce reliance on Likert-type self-reports at the early stages of the hiring process, given the ease with which applicants can artificially inflate—or fake—their scores (Morgeson et al. 2007).

Faking in the application process is problematic and occurs in both self-reports and interviews (Buehl et al. 2019; Canagasuriam and Roulin 2021; Van Iddekinge et al. 2005). Such response distortions artificially inflate scores and decrease the construct validity of self-reports (Arthur et al. 2021) and interviews (Van Iddekinge et al. 2005). Further, faking is difficult to detect reliably, and correcting self-reports for faking raises fairness and validity issues. Although these limitations were one motivation for developing alternatives to self-reports such as AI scored interviews, we know little about how AI scored interviews are affected by applicant faking (Hickman et al. 2022). Given that the AI applies a consistent decision rule

Summary

- Automatically scored asynchronous interviews are increasingly being used to replace traditional assessments, such as self-report personality tests, in high-stakes selection
- However, we know little about the extent to which applicants can inflate their scores on interviews scored by machine learning
- Participants in the study completed an interview at two time points-once under honest instructions and once under applicant-like instructions
- Machine learning scores were no more susceptible to response inflation than human ratings, and all interview scores were much less susceptible to response inflation than self-reports
- These findings suggest that automatically scored asynchronous interviews hold promise for overcoming the faking problem with self-report personality inventories

to evaluate interviews but has no awareness about the reasonableness or consistency of responses—when no system has been trained to recognize response distortions—AI scoring may be particularly prone to applicant faking and impression management (IM; Holtrop et al. 2022). Further, the mechanisms of score inflation may differ from traditional interviews.

Thus, the present study investigates faking and score inflation for AI scored interviews. AI scored interviews are asynchronous, technology-mediated interviews, and thus, the magnitude of score inflation and the mechanisms of score inflation may differ in them compared to face-to-face interviews (Melchers et al. 2020). Specifically, previous research suggests that cognitive ability, ability to identify criteria (ATIC; Kleinmann et al. 2011), and IM explain how interviewees increase their interview scores. Overall, we aim to replicate and extend prior research on faking by expanding the investigation to include AI scored asynchronous interviews.

To do so, participants completed a mock asynchronous interview and self-reports under both honest- and applicant-like conditions. The interviews were then scored by humans (using both a less structured and highly structured approach) and AI models. Then, we compared the extent of score inflation and effects on construct validity across self-reports and the different interview scoring methods. Next, we examined whether verbal ability, ATIC, and IM predict the extent of score inflation for self-reports and the three interview scoring approaches when moving from honest to applicant-like experimental conditions. In doing so, our study expands investigations of applicant faking to AI scored interviews and asynchronous interviews while constructively replicating prior research on the mechanisms that explain interview score inflation (e.g., Buehl et al. 2019; Buehl and Melchers 2017).

1 | AI Interview Scores and Score Inflation

In recent years, researchers and practitioners have grown interested in using AI to score open-ended assessments (Campion et al. 2016; Thompson et al. 2023), including employment interviews (Hickman et al. 2022; Liff et al. 2024). AI models for scoring open ended assessments are generally designed according to the following best practices. First, a large sample of interviews is collected—whether specifically for the purpose of training an AI model, or archival data collected for other purposes. Second, multiple human evaluators review the interviews and rate interviewee performance. Third, the interviews are transcribed using AI models. Fourth, NLP methods are applied to the transcripts to quantify interviewee responses. Fifth, a supervised ML model is trained using the NLP variables as predictors of the human interview ratings. Our research focuses on models trained to replicate human ratings on behaviorally anchored rating scales (BARS) of a question response (e.g., Liff et al. 2024), although other researchers have also examined scoring constructs based on the interviewee's responses across the entire interview (e.g., Hickman et al. 2022; Jayaratne and Jayatilleke 2020).

Considerable evidence is now available that such automatic interview scores can converge highly with human interviewer ratings (Hickman et al. 2022; Liff et al. 2024). The promise of using supervised ML models for scoring is that the models are completely consistent in their evaluations. The algorithm never grows tired or inattentive, never has a bad day, and will give an identical score to the same interview whether evaluated minutes, days, or years apart.[1] Notably, this does not mean that the AI scores are necessarily fair and unbiased, but it does provide perfect intrarater reliability. Of course, updating or changing the AI model would alter the scores assigned.

Although AI scored interviews can serve as alternatives to self-reports, it is unknown whether such AI models can be faked (Hickman et al. 2022). Self-reports exhibit considerable score inflation between honest and applicant-like conditions (and between incumbents and applicants), which creates fairness and validity concerns when using them for selection. Human scored interviews exhibit less score inflation than self-reports (Van Iddekinge et al. 2005). Multiple studies have demonstrated that interviewees can nonetheless improve their interview scores when moving from honest to applicant-like conditions in face-to-face interviews (e.g., Buehl et al. 2019; Buehl and Melchers 2017). Indeed, sometimes the overall score inflation can be more than a standard deviation (e.g., Buehl et al. 2019 found overall interview scores increased $d = 1.33$ for an interview consisting of 20 questions). However, we do not know how score inflation in AI scored interviews compares to self-reports or human-scored interviews.

Researchers have also suggested that the extent of score inflation should be compared across interviews of varying structure (Melchers et al. 2020). We address this by comparing score inflation across different levels of scoring structure applied to the same interview responses. We refer to "less structured human ratings" as ratings completed by raters who underwent no frame of reference training (e.g., Roch et al. 2012) and who did not utilize BARS, and we refer to "structured human ratings" as ratings completed by raters who underwent frame of reference training and used BARS (in line with categories of interview structure described by Huffcutt et al. 2013).

The AI models in the present study were trained on structured human ratings, with a separate model for each interview question. Thus, we expect that the AI will exhibit score inflation similar to the structured human ratings. Further, given that interview ratings tend to exhibit less score inflation than self-reports (Van Iddekinge et al. 2005), we also expect that the AI scores will exhibit less score inflation than self-reports.

Less structured human ratings may be more prone to score inflation because they can be affected by factors irrelevant to interview performance (e.g., paraverbal and nonverbal behaviors). However, less structured human ratings' score inflation may be attenuated by their relatively lower reliability (Huffcutt et al. 2013; Van Iddekinge et al. 2005), and this lower reliability also partially explains why they tend to exhibit lower validity than more structured ratings. Thus, we make no prediction about the relative score inflation of less structured scores compared to either AI scores or more structured human ratings.

Hypothesis 1a-1b. *AI interview scores are less susceptible to score inflation than* (*a*) *self-reported surveys but as susceptible to score inflation as* (*b*) *structured interviewer ratings.*

The structured human ratings in our study were made on the same BARS that was used when generating training data for the AI model. Thus, we expect that in both the honest- and applicant-like conditions, the AI model scores will converge more strongly with the structured human interviewer ratings than the less structured interviewer ratings.

Hypothesis 2. *In the honest- and applicant-like conditions, the AI interview scores will converge more strongly with structured interviewer ratings than with less structured interviewer ratings.*

2 | Mechanisms of Score Inflation

Prior research has sought to understand the characteristics that make interviewees effective at faking and providing socially desirable responses in interviews. We seek to replicate these findings regarding human interview ratings, and we extend them by also investigating them for asynchronous, automatically scored interviews. In terms of individual differences, interviewee ability has emerged as a powerful predictor of both interview performance and the capacity to fake (Buehl et al. 2019; Buehl and Melchers 2017; König et al. 2007; Melchers et al. 2009). We argue that verbal ability and ability to identify criteria (ATIC) are important for understanding faking effectiveness.

Verbal ability is one of the most important components of general mental ability (GMA), given that several GMA tests oversample from verbal ability (e.g., the Wonderlic), and that verbal ability explains much of the relationship between GMA and job performance (Lang and Kell 2019; Schneider and Newman 2015). When components of GMA are measured separately, verbal ability tends to predict interview performance while other abilities do not (e.g., König et al. 2007; Melchers et al. 2009). Further, some employment interviews are considered to be verbally-administered ability tests (Hunter and Hirsh 1987), suggesting that verbal ability is the ability most directly related to interview performance.

Prior research has found that cognitive ability relates to interview regression-adjusted difference scores (RADS), or the part of an interviewee's score in an applicant-like condition that cannot be explained by their scores in the honest condition (Buehl et al. 2019; Buehl and Melchers 2017). These prior studies measured cognitive ability using the Wonderlic, which oversamples from verbal ability, finding correlations with interview RADS of $r = 0.18$ & 0.19. Thus, we expect that verbal ability will relate positively with interview RADS. Further, we expect this pattern to occur across AI and human interview scoring methods.

Hypothesis 3. *Interviewee verbal ability correlates positively with interview score improvement when moving from honest- to applicant-like condition.*

Ability to identify criteria (ATIC) is an individual difference in how accurately people can identify the constructs measured in interviews and other evaluative situations (Kleinmann et al. 2011; König et al. 2007). The expectation is that people who know what is being sought by evaluators will be better able to enact desired behaviors. Indeed, ATIC correlates positively with interview and assessment center performance (Griffin 2014; Ingold et al. 2015; König et al. 2007; Melchers et al. 2009). In this way, ATIC is also thought to be an important component of successful faking. When interviewees can identify the construct being assessed, it should improve their ability to provide descriptions of relevant, positive behaviors. Prior research has also found that ATIC is related to score improvement when moving from honest- to applicant-like conditions (Buehl et al. 2019). Thus, we expect that the same trend will occur here, regardless of the interview scoring method.

Hypothesis 4. *Interviewee ATIC correlates positively with interview score improvement when moving from honest- to applicant-like condition.*

In terms of specific behaviors, IM behaviors are critically important in interviews (Bourdage et al. 2018). Impression management involves self-presentation behaviors that people use with the goal of improving the impressions they make on other people (Leary and Kowalski 1990). Effective IM use can affect interview outcomes (Barrick et al. 2009).

In recent decades, researchers have delineated two broad classes of IM behaviors: honest and deceptive (Bourdage et al. 2018; Higgins and Judge 2004). Honest IM tactics include self-promotion, ingratiation, defensive, and nonverbal IM. Self-promotion involves emphasizing the job relevant knowledge, skills, and experiences that they hold (Levashina et al. 2014). Ingratiation involves "other-focused" behaviors aimed at making the interviewer like them more, such as laughing at their jokes and complimenting them. Given that the present study examines asynchronous interviews, we did not measure or consider ingratiation IM tactics. Defensive IM involves providing explanations and justifications for why negative events occurred and is a more reactive IM method. Nonverbal IM is slightly different, in that it is more of a self-presentation tactic rather

than one that involves sharing useful information. Nonverbal IM should have no influence on AI ratings or structured human ratings, given that those evaluations focus exclusively on the content of the response.

Deceptive IM tactics include slight image creation, extensive image creation, and image protection (Levashina and Campion 2007). Slight image creation involves slightly embellishing the truth, such as taking more credit for an event than deserved, while extensive image creation involves a greater level of fabrication, such as taking credit for an event when not being involved in its design or execution. Image protection involves lies by omission, such as not mentioning negative events that occurred or withholding the full truth in the interview.

Honest IM tactics tend to exhibit small, positive effects on interview ratings (Peck and Levashina 2017). Since honest IM involves better "selling" your past experiences, this is unsurprising. However, deceptive IM tends to be unrelated to interview performance (Ho et al. 2021). Although the reason is unclear, it may be that people who need to make up stories are not necessarily the best storytellers, which reduces the effectiveness of deceptive IM. Thus, we expect that increased usage of honest IM—but not deceptive or nonverbal IM—will increase interview scores in the applicant-like condition. Again, we expect this pattern to be consistent across AI scored, structured, and less structured human rating methods.

Hypothesis 5. *Increased use of honest impression management correlates positively with interview score improvement when moving from honest- to applicant-like condition.*

There is major interest in detecting socially desirable responding and faking in interviews (e.g., Auer 2018; Roulin and Powell 2018). However, accuracy is often quite poor in these approaches—at least partly because there are few reliable behavioral cues of deception (Vrij et al. 2010). Melchers et al. (2020) recently suggested that it is important to explore the differences in behavior that occur when interviewees respond honestly versus when they are motivated to engage in impression management and faking. Thus, we apply natural language processing (NLP) to explore the differences in interview responses across our experimental conditions.

Research Question 1. *How do interviewee responses change from the honest- to applicant-like condition?*

3 | Methods

The study was approved by the University of Pennsylvania IRB (850389). The OSF repository includes the experimental materials, data, and analytical scripts: https://osf.io/dhwgm/.

3.1 | Participants and Procedure

At Time 1 (T1) participants were randomly assigned to either an honest condition or an applicant-like condition (in line with prior research; e.g., Buehl et al. 2019). In the honest condition, participants were instructed to respond in a way that reflected their true perceptions of their personality, behavior, and past performance, and they were told that they would be randomly entered to receive a $50 bonus. In the applicant-like (faking) condition, participants were provided with a hiring scenario in which they were instructed to respond as if they were applying for a Project Manager role at a fictitious organization. In this condition, they were also told that the top five scoring interviewees would receive a $50 bonus. Thus, in addition to the regular study payments (described below), participants were eligible to win one or two of the 10 $50 bonuses if they completed both interview time points.

At Time 2 (T2) participants were assigned the opposite condition they completed at T1.

Specifically, those who were assigned to the honest condition at T1 were assigned to the applicant-like condition at T2 and vice-versa. Participants were eligible to begin the T2 interview process 10 days after completing T1 and had 14 days to complete T2.

At both time points, participants completed a mock interview consisting of five questions. The questions were designed to measure the competencies of adaptability, initiative, composure, dependability, and problem-solving (Borman et al. 1999). Participants had as much time as they desired to prepare before answering. Their responses were limited to a maximum of 3 min, and they were allowed to re-record their answer as many times as they liked. Afterwards, participants completed self-reports of conceptually similar constructs, their impression management tactics in the interview, demographics, and completed a treatment check. Finally, participants were debriefed, thanked for their participation, and received their bonuses if eligible. The entire study was conducted in an online platform for conducting asynchronous video interviews.

Based on a priori power analysis (see power analysis calculator in OSF repository), 175 participants provided 80% power to detect a Cohen's d effect size of 0.30 at $p < 0.05$ across the experimental conditions. Thus, we aimed to recruit 200 participants at T1, expecting that 175 participants would complete T2. At T1, we recruited 203 participants from Prolific. A total of 185 participants (91.1%) completed all study procedures at T1 and T2. These participants were compensated $22.50 for their participation. Of these participants, 152 (82.2%) correctly answered the treatment check at both time points. The majority of participants (43.4%) were aged 24–34 and 35–44 (28.3%), and 46.7% were female, 51.3% were male, and 2.0% were other/nonbinary/did not want to specify. In terms of race/ethnicity, 65.3% of participants reported being White, 13.3% Asian, 12.0% two or more races/ethnicities, and 5.3% Black. The majority of participants (95.4%) were employed full-time and had a Bachelor's degree (56.6%) or a Master's degree or above (41.4%).

Approximately 1 year later, we invited all participants who passed both treatment checks to complete an additional survey for $6. The additional survey included the verbal ability and ATIC measures. $N = 105$ participants completed this follow up survey, and 104 of them provided usable ATIC responses.

3.2 | Measures

In addition to other measures in the study, all participants were evaluated with four methods on adaptability, composure, dependability, initiative, and problem solving. These constructs were selected for their content relevance to project manager roles following a job analysis conducted by HireVue to advise clients on which constructs to assess for such roles.

3.2.1 | Self-Reports

All self-reports were measured on five-point Likert scales ranging from strongly disagree to strongly agree. Scales were selected that aligned closely with the definitions of the constructs provided in the BARS used in the study.

3.2.1.1 | Adaptability. Adaptability was measured using the International Personality Item Pool (IPIP; Goldberg et al. 2006) adaptation of the Six Factor Personality Questionnaire (6FPQ)'s subscale (Jackson et al. 2000). The adaptability scale has a total of eight items, including, "Can stand criticism", and "Don't tolerate critics" (reverse scored).

3.2.1.2 | Composure. Composure was measured using the IPIP calmness scale (Goldberg et al. 2006) that corresponds to a facet in the Hogan Personality Inventory (HPI) (Hogan et al. 1992). The scale includes six items, such as, "Remain calm under pressure", and "Get overwhelmed by emotions" (reverse scored).

3.2.1.3 | Dependability. Dependability was measured using the responsibility facet scale from the Big Five Inventory – 2 (BFI-2), a revision to the original BFI (Soto and John 2017). Responsibility is a facet of conscientiousness in the BFI-2. The scale includes four items, including, "Is dependable, steady", and "Can be somewhat careless" (reverse scored).

3.2.1.4 | Initiative. Initiative was measured using the productiveness facet scale from the BFI-2 (Soto and John 2017). Productiveness is a facet of conscientiousness in the BFI-2. The scale includes four items, such as, "Is efficient, gets things done", and "Tends to be lazy" (reverse scored).

3.2.1.5 | Problem Solving. Problem solving was measured using the IPIP science ability scale (Goldberg et al. 2006) that corresponds to a facet in the HPI (Hogan et al. 1992). The scale includes six items, including, "Like to solve complex problems", and "Dislike learning" (reverse scored).

3.2.2 | Interview and Interview Ratings

The interview comprised five questions, each designed to assess one of the five constructs (i.e., adaptability, composure, dependability, initiative, and problem solving). We cannot divulge the interview questions because they are currently being used in high-stakes interviews. Participants were instructed to respond for a minimum of 45 s, and the recording stopped either when the participant stopped the recording or after 3 min, whichever occurred first. Each interview question response was then scored by an AI, structured human ratings, and less structured human ratings.

3.2.2.1 | AI Interview Ratings. The question-level responses were scored by a set of ML models that are used to automatically score interviews in high-stakes settings. More details about the general development process and validation evidence are provided in Liff et al. (2024). The AI system applies natural language processing (NLP) to the interview transcripts, and each model was trained on thousands of observations from applicants across many types of jobs to predict structured human ratings. The human raters underwent frame-of-reference training and rated question-level performance on behaviorally anchored rating scales. Separate models are trained for each construct, using answers to and ratings of several structured interview questions designed to measure that construct.

The NLP uses a combination of fine-tuned RoBERTa embeddings and *n*-grams to operationalize interviewee verbal behavior. These predictors are used in a custom ridge regression model that also penalizes regression weights of predictors that contribute to adverse impact (Rottman et al. 2023). No paraverbal or nonverbal behaviors are used by the system, and the data in the present study was not used to train the models. These AI scores have demonstrated strong convergent correlations with structured human ratings ($rs \geq 0.65$), good test-retest reliability ($r_{tt} \geq 0.65$), and criterion-related validity for both objective and subjective measures of future job performance ($rs \geq 0.23$; Liff et al. 2024).

3.2.2.2 | Structured Human Interview Ratings. A pool of eight graduate and undergraduate research assistants underwent 1–2 h of frame-of-reference training for each individual question and completed ratings of that question before moving on to the next. The frame-of-reference training consisted of defining the focal construct, reviewing the BARS and the behavioral anchors, then completing practice ratings and discussing them as a group. Then, three to four of them reviewed each recording and rated interviewee performance on a BARS ranging from 1 (novice) to 5 (expert). The BARS included anchors for scores of one, three, and five for each of four to five subdimensions that evaluators considered when making their final, overall rating.

3.2.2.3 | Less Structured Human Interview Ratings. A similar procedure was used to collect a set of less structured interview ratings, except that these undergraduate research assistants did not undergo frame-of-reference training and received only the construct definition portion of the BARS (i.e., did not receive the behavioral anchors). From a pool of 12 undergraduate research assistants, 3–4 rated each question response.

3.2.3 | Faking Effectiveness

To operationalize faking effectiveness, we used the regression-adjusted difference scores (RADS) between the honest and

faking condition (Burns and Christiansen 2011; as used in Buehl and Melchers 2017). The RADS are calculated by regressing the applicant-like condition scores on the honest condition scores, then saving the residuals. The residuals represent the part of the applicant-like scores that could not be explained by the honest condition scores. For concision in the manuscript, we focus on the overall RADS for each scoring method (i.e., based on the average score on the five constructs), but we report the RADS at the construct level in the online supplement.

3.2.4 | Impression Management

Impression management tactics were measured using the short impression management scale (Bourdage et al. 2018), adapted to a one-way/asynchronous interview context (i.e., removing any mention of an interviewer). Out of the three total honest IM tactics, subscales for self-promotion and defensive IM were included. Of the three deceptive subscales, all three (deceptive slight image creation, deceptive extensive image creation, and deceptive image protection) were measured. Additionally, nonverbal IM was measured using a scale adapted from Tsai et al. (2005) for the asynchronous interview context. Each subscale, other than nonverbal, consists of four items each. It should be noted that while the nonverbal IM scale consists of three items, two of the items were inadvertently presented as a single item during the setup of the study, thus reducing the scale to two items. To operationalize impression management when testing Hypothesis 5, we use IM RADS, which represents the impression management behaviors in the applicant-like condition that could not be explained by impression management behaviors in the honest condition.

3.2.5 | Treatment Checks

At both time points, after completing the interview and the self-reports of the five competencies (but before other measures), participants were asked, "What instructions did you receive for the study today?" The response options included, "To respond honestly," "To respond as an applicant for a sales role," and "To respond as an applicant for a project manager role." Participants who failed the treatment check at one or both time points were excluded from our final analyses.

3.2.6 | Attention Checks

At both time points, an attention check item was included in both the self-reports of the five competencies and in the IM scale items. These included explicit attention checks (e.g., "I am someone who selects 'Strongly agree' if I am paying attention to this survey.") and items asking about behaviors that were definitely not engaged in during the interview (e.g., "During the interview, I ate ghost peppers."). Across the two time points, no participants who passed both treatment checks failed more than one attention check, so no additional data was excluded due to these items.

3.2.7 | Verbal Ability

Verbal ability was measured with a custom, 19-question multiple choice test designed to mimic questions on the verbal portion of the Graduate Record Examinations (GRE). Validation details are provided in Hickman et al. (2025). The test exhibits convergent and discriminant correlations that suggest it measures verbal ability.

3.2.8 | Ability to Identify Criteria (ATIC)

Ability to identify criteria was measured with methods consistent with prior studies (e.g., Buehl et al. 2019). Specifically, in the follow up study, participants were presented with each of the five interview questions. Two open-ended response boxes were provided. The first instructed participants to report which dimension(s) they believed were assessed with the interview question, and the second asked them to give behavioral examples for the targeted dimension(s).

After completing the survey, two trained raters examined the participants' responses for each interview question and rated how well these matched the targeted dimension using a scale from 0 (no fit) to 3 (fits completely). Before discussion, raters agreed 60.8% of the time and their ratings correlated $r = 0.70$, with 96.6% of disagreements being by one point. Raters then discussed all disagreements and adjusted their ratings, as they considered appropriate. After discussion, the raters agreed 92.9% of the time and the two sets of ratings correlated $r = 0.96$. Each participant's final ATIC score was calculated as the average of the two raters' scores across the five questions.

3.2.9 | Measuring Behavioral Changes

To measure differences in interviewee responses across the two conditions, we applied multiple NLP techniques. First, we applied Linguistic Inquiry and Word Count (LIWC; Pennebaker et al. 2015) to the transcripts and compared means across the experimental conditions. LIWC measures a series of psycholinguistic dictionaries and captures changes in the content of responses in the interview. Second, we used the quanteda R package to calculate readability statistics from the transcripts. Specifically, we calculated Flesch's reading ease score and FORCAST (Caylor and Sticht 1973). The readability statistics capture changes in the style of interview responses.

4 | Results

Table 1 reports the descriptive statistics of the study variables. As a manipulation check, we first examine whether the self-report, interview, and impression management scores were higher in the applicant-like than honest condition. The mean scores across the five constructs increased within-subjects from the honest to applicant-like condition (ds = 0.62, 0.14, 0.22, and 0.17, respectively, for self-reports, AI interview scores, structured human ratings, and less structured human ratings), although the interview score increases were substantially

TABLE 1 | *Descriptive Statistics of Study Variables.*

Variable	Honest			Applicant-like			Overall
	Mean	SD	Reliability	Mean	SD	Reliability	Test-retest (r_{tt})
Gender	0.48	0.50					
Age	2.74	1.11					
Self-reports							
Adaptability	4.06	0.46	0.65	4.33**	0.46	0.76	0.50
Composure	4.17	0.65	0.88	4.47**	0.59	0.91	0.48
Dependability	4.38	0.59	0.78	4.65**	0.45	0.75	0.58
Initiative	4.21	0.68	0.78	4.52**	0.51	0.73	0.57
Problem solving	4.22	0.44	0.63	4.35**	0.45	0.74	0.48
AI interview ratings							
Adaptability	3.21	1.01		3.36	1.05		0.47
Composure	3.64	0.86		3.78*	0.85		0.55
Dependability	3.18	0.81		3.33	0.81		0.37
Initiative	3.25	0.86		3.26	0.92		0.47
Problem solving	3.28	0.78		3.29	0.82		0.38
Structured human ratings							
Adaptability	3.26	0.62	0.58	3.39*	0.64	0.59	0.36
Composure	2.97	0.63	0.69	3.13**	0.64	0.69	0.43
Dependability	2.85	0.67	0.68	2.94	0.64	0.66	0.38
Initiative	2.94	0.63	0.66	3.04*	0.65	0.68	0.55
Problem solving	3.04	0.54	0.59	3.08	0.56	0.65	0.41
Less structured human ratings							
Adaptability	3.32	0.67	0.64	3.29	0.66	0.65	0.50
Composure	3.36	0.75	0.61	3.40	0.65	0.59	0.40
Dependability	3.30	0.68	0.59	3.41	0.67	0.55	0.46
Initiative	3.23	0.74	0.57	3.36	0.77	0.64	0.36
Problem solving	3.20	0.67	0.63	3.37**	0.63	0.60	0.36
Impression management							
HSP	3.33	0.79	0.72	3.76**	0.75	0.75	0.51
HD	2.45	0.86	0.71	2.52	0.79	0.61	0.52
DSIC	1.32	0.55	0.80	1.53**	0.72	0.85	0.47
DEIC	1.25	0.51	0.70	1.40*	0.76	0.86	0.37
DIP	1.65	0.72	0.74	1.90**	0.88	0.73	0.42
Nonverbal	2.79	0.80	0.55	2.89	0.89	0.58	0.70
Verbal ability	20.49	7.47					
ATIC	1.43	0.47					

Note: ATIC = ability to identify criteria. HSP = honest self-promotion. HD = honest defensive. DSIC = deceptive slight image creation. DEIC = deceptive extensive image creation. DIP = deceptive image protection. $N = 152$ for all variables except verbal ability and ATIC, where $N = 105$ and 104, respectively. For self-reports, verbal ability test, and ATIC, reliability is Cronbach's α. For interviewer ratings, reliability is $G(q, k)$ (Putka et al. 2008).
*indicates $p < 0.05$ comparing honest and applicant-like conditions and
**indicates $p < 0.01$. Age was measured in six categories: Under 25, 25–34, 35–44, 45–54, 55–64, 65–74, and over 75.

smaller than those observed for in-person interviews between-subjects by Van Iddekinge et al. (2005) or within-subjects by Buehl et al. (2019). Additionally, four impression management behaviors (honest self-promotion, deceptive-slight image creation, deceptive-extensive image creation, and deceptive-image protection) increased significantly from the hones tto applicant-like condition. Together, these results suggest that the manipulation was effective at getting participants to make a more applicant-like impression.

Table 2 reports the correlation matrix for: verbal ability; ATIC; overall scores (i.e., the average of the five construct scores) on

TABLE 2 | *Correlations Among the Study Variables.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Verbal ability																					
2. Ability to identify criteria	−0.05																				
3. Self-reports (honest)	0.05	−0.12																			
4. Self-reports (applicant)	0.23	−0.03	0.57																		
5. AI ratings (honest)	0.21	0.21	0.10	0.14																	
6. AI ratings (applicant)	0.13	0.15	−0.03	0.06	0.66																
7. Structured ratings (honest)	0.22	0.16	0.12	0.16	0.75	0.63															
8. Structured ratings (applicant)	0.20	0.13	0.00	0.13	0.59	0.76	0.71														
9. Less structured ratings (honest)	0.24	0.19	0.08	0.13	0.71	0.60	0.80	0.70													
10. Less structured ratings (applicant)	0.22	0.16	−0.03	0.09	0.56	0.74	0.65	0.78	0.68												
Honest condition IM																					
11. Honest self promotion	−0.12	0.09	0.12	0.09	0.07	0.08	0.06	0.01	0.01	0.05											
12. Honest defensive	−0.11	−0.18	0.19	0.03	−0.08	−0.08	−0.13	−0.10	−0.17	−0.16	0.27										
13. DSIC	−0.19	0.09	−0.40	−0.24	−0.18	−0.07	−0.10	−0.05	−0.11	−0.03	0.11	0.08									
14. DEIC	−0.13	0.10	−0.30	−0.04	−0.09	0.01	0.00	0.02	0.03	0.08	−0.03	−0.05	0.54								
15. Deceptive image protection	−0.13	0.23	−0.30	−0.12	−0.01	−0.03	0.04	0.04	0.00	0.01	0.11	−0.01	0.52	0.34							
16. Nonverbal	−0.05	−0.02	0.31	0.29	0.04	−0.03	0.08	0.03	0.01	−0.01	0.31	0.25	−0.05	−0.03	−0.03						
Applicant condition IM																					
17. Honest self promotion	0.06	−0.09	0.14	0.27	0.03	0.11	0.05	0.07	0.02	0.10	0.51	0.15	−0.04	−0.03	−0.06	0.27					
18. Honest defensive	−0.02	−0.08	0.19	0.13	−0.06	0.06	−0.03	−0.05	−0.17	−0.01	0.08	0.52	0.09	0.06	−0.01	0.14	0.28				
19. DSIC	−0.04	0.21	−0.27	−0.21	−0.11	−0.03	−0.05	0.03	−0.08	0.01	−0.10	0.00	0.47	0.30	0.39	−0.10	−0.08	0.03			
20. DEIC	−0.15	0.20	−0.11	−0.05	−0.05	−0.10	−0.03	−0.07	−0.07	−0.06	−0.14	0.09	0.28	0.37	0.25	0.02	−0.10	0.10	0.67		
21. Deceptive image protection	−0.02	0.15	−0.19	−0.09	0.08	0.04	0.05	0.04	−0.02	−0.02	0.06	0.06	0.18	0.10	0.42	−0.04	0.06	−0.06	0.46	0.42	
22. Nonverbal	0.05	0.08	0.20	0.24	0.07	0.05	0.13	0.13	0.05	0.08	0.15	0.12	0.02	0.04	0.07	0.70	0.27	0.25	0.03	0.14	0.02

Note: ATIC = ability to identify criteria. Verbal ability $N = 105$; ATIC $N = 104$; $N = 152$ for all remaining variables. For verbal ability and ATIC, correlations $|r| > 0.19$ have $p < 0.05$ (two-tailed). For remaining variables, $|r| > 0.16$ have $p < 0.05$ (two-tailed). The self-reports, AI ratings, structured human ratings, and less structured human ratings are the average score received across the five constructs in that condition-scoring method pair. DSIC = deceptive slight image creation. DEIC = deceptive extensive image creation.

self-reports, AI interview scores, structured human interview ratings, and less structured human interview ratings; and self-reported impression management tactics.

4.1 | Within-Subject Results Across Timepoints

Table 3 reports the Cohen's *d*s, Cohen's *d*s corrected for unreliability, and confidence intervals of the corrected Cohen's *d*s for the increase in scores moving from honest to applicant-like conditions for self-reports, AI interview scores, structured human interview ratings, and less structured human interview ratings. As can be seen, for all five constructs, self-report scores increased significantly (d_c ranging from 0.35 to 0.69). AI interview scores increased significantly for adaptability ($d_c = 0.19$), composure ($d_c = 0.15$), dependability ($d_c = 0.18$), and overall scores ($d = 0.14$). Structured human scores increased significantly for adaptability ($d_c = 0.25$), composure ($d_c = 0.32$), initiative ($d_c = 0.20$), and overall ($d = 0.22$). Less structured human scores increased significantly for dependability ($d_c = 0.21$), initiative ($d_c = 0.22$), problem solving ($d_c = 0.32$) and overall ($d = 0.17$).

TABLE 3 | *Effect size differences between applicant-like and honest conditions.*

Adaptability	*d*	d_c	d_c 95% CI
Self-reports	0.59**	0.69**	[0.51, 0.86]
AI interview ratings	0.15	0.19*	[0.02, 0.35]
Structured human ratings	0.20*	0.25**	[0.06, 0.43]
Less structured human ratings	−0.04	−0.05	[−0.21, 0.11]
Composure			
Self-reports	0.49**	0.55**	[0.40, 0.71]
AI interview ratings	0.15*	0.18*	[0.03, 0.34]
Structured human ratings	0.26**	0.32**	[0.14, 0.49]
Less structured human ratings	0.06	0.08	[−0.09, 0.26]
Dependability			
Self-reports	0.50**	0.53**	[0.37, 0.68]
AI interview ratings	0.18	0.22*	[0.04, 0.40]
Structured human ratings	0.13	0.16	[−0.02, 0.34]
Less structured human ratings	0.16	0.21*	[0.04, 0.38]
Initiative			
Self-reports	0.50**	0.55**	[0.38, 0.72]
AI interview ratings	0.01	0.01	[−0.16, 0.17]
Structured human ratings	0.16*	0.20*	[0.05, 0.35]
Less structured human ratings	0.17	0.22*	[0.03, 0.40]
Problem solving			
Self-reports	0.29**	0.35**	[0.18, 0.51]
AI interview ratings	0.02	0.02	[−0.16, 0.19]
Structured human ratings	0.07	0.09	[−0.08, 0.26]
Less structured human ratings	0.25**	0.32**	[0.14, 0.50]
Overall			
Self-reports	0.62**		
AI interview ratings	0.14*		
Structured human ratings	0.22**		
Less structured human ratings	0.17*		

**$p < 0.01$; *$p < 0.01$. d_c is *d* corrected for unreliability, based on the formula in Bobko et al. (2001). For self-reports and human interview scores, reliability is reported in Table 1. For the ML interview scores, reliability is reported in Liff et al. (2024).

To answer Hypothesis 1 (i.e., whether AI interview scores are less susceptible to score inflation than self-reports but as susceptible as structured interviewer ratings), we compared the 95% confidence intervals for *d* and d_c in Table 3 across methods. Self-reports exhibited a larger increase moving from honest to applicant-like conditions than the AI interview scores for the overall scores (i.e., average of the five constructs), adaptability, composure, and initiative, while the effects were in the hypothesized direction but nonsignificant for dependability and problem solving. In all cases, the confidence intervals for the AI interview scores and structured human scores overlapped. Thus, Hypothesis 1a was largely supported, and Hypothesis 1b was fully supported.

To further examine Hypothesis 1, we also consider the test-retest reliability of the assessment methods across the experimental conditions. Test-retest reliability across conditions will be lower when (1) there is more variance in the extent to which participants are able to improve their scores in the applicant-like condition and (2) when a measure is less reliable (Gordon and Gross 1978). As reported in Table 1, test-retest reliability averaged $\bar{r}_{tt} = 0.52$ for self-reports, 0.45 for AI interview scores, 0.43 for structured human interview scores, and 0.42 for less structured human interview scores. However, these differences are not significant.

Next, we examined the construct validity of the scores between the two conditions, given that method variance is another indicator of faking (Burns and Christiansen 2011). Heterotrait-monomethod (HTMM; discriminant) correlations tended to increase in the applicant-like condition, as reported in Table 4. Self-report scores' HTMM correlations increased from $\bar{r} = 0.43$ in the honest condition to $\bar{r} = 0.57$ in the applicant-like condition. The AI interview scores' HTMM correlations increased from $\bar{r} = 0.38$ to $\bar{r} = 0.44$ in the applicant-like condition. Structured human scores' HTMM correlations did not increase ($\bar{r}$s = 0.44 & 0.42, respectively, in the honest and applicant-like conditions), but less structured human scores' HTMM correlations increased ($\bar{r}$s = 0.36 & 0.42, respectively, in the honest and applicant-like conditions) at a magnitude similar to the AI scores. The increase in HTMM correlations across conditions was only substantial for self-reports, while it was similar for AI and human interview ratings, further supporting Hypotheses 1a and 1b.

Hypothesis 2 proposed that AI interview scores would converge more strongly with structured interviewer ratings than less structured ones across both conditions. In the applicant-like

TABLE 4 | *Multi-trait multi-method matrix of self-reports and interview scores (Honest condition below diagonal, applicant-like above).*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Self-reports																				
1. Adaptability	—	0.65	0.53	0.56	0.46	0.06	0.06	0.11	0.13	0.09	0.09	0.08	0.20	0.17	0.01	0.05	0.12	0.09	0.12	0.09
2. Composure	0.54	—	0.65	0.59	0.53	0.00	-.01	0.10	0.04	0.04	0.11	0.08	0.12	0.14	0.05	0.08	0.10	0.07	0.08	0.05
3. Dependability	0.27	0.42	—	0.68	0.59	0.04	0.00	0.09	0.10	0.06	0.14	0.07	0.13	0.04	0.06	0.06	0.04	0.07	0.10	0.11
4. Initiative	0.39	0.42	0.74	—	0.49	0.04	0.02	0.06	0.04	0.03	0.16	0.10	0.08	0.05	0.06	0.04	0.04	−0.02	0.07	0.08
5. Problem solving	0.31	0.36	0.45	0.40	—	0.00	−0.05	0.00	−0.07	−0.01	−0.01	−0.07	−0.04	0.04	−0.03	−0.07	−0.05	−0.02	−0.01	0.00
AI interview ratings																				
6. Adaptability	0.09	0.07	0.08	0.09	0.03	—	0.50	0.36	0.55	0.48	0.48	0.40	0.32	0.44	0.31	0.56	0.28	0.36	0.45	0.33
7. Composure	0.03	0.04	0.04	−0.01	−0.02	0.44	—	0.53	0.55	0.39	0.40	0.60	0.51	0.52	0.43	0.46	0.34	0.47	0.49	0.43
8. Dependability	0.11	0.04	0.03	0.05	0.03	0.42	0.48	—	0.37	0.32	0.15	0.40	0.61	0.48	0.34	0.33	0.42	0.59	0.40	0.40
9. Initiative	0.06	0.02	0.09	0.15	0.09	0.46	0.45	0.38	—	0.31	0.43	0.47	0.39	0.54	0.32	0.46	0.35	0.40	0.55	0.41
10. Problem solving	0.07	0.03	0.05	0.06	0.01	0.28	0.29	0.33	0.26	—	0.34	0.30	0.30	0.38	0.50	0.38	0.24	0.31	0.27	0.33
Structured human ratings																				
11. Adaptability	0.11	0.07	−0.01	−0.05	−0.01	0.44	0.29	0.27	0.26	0.35	—	0.42	0.29	0.38	0.41	0.63	0.32	0.30	0.31	0.23
12. Composure	0.15	0.06	0.08	0.05	0.03	0.37	0.53	0.43	0.43	0.29	0.36	—	0.45	0.44	0.45	0.51	0.55	0.45	0.42	0.44
13. Dependability	0.10	0.00	0.02	0.08	0.04	0.38	0.38	0.67	0.39	0.31	0.41	0.48	—	0.55	0.32	0.42	0.29	0.65	0.41	0.26
14. Initiative	0.15	0.12	0.06	0.09	0.10	0.41	0.37	0.47	0.50	0.36	0.44	0.43	0.49	—	0.49	0.46	0.34	0.47	0.62	0.43
15. Problem solving	0.12	0.14	0.09	0.08	−0.03	0.38	0.38	0.35	0.35	0.48	0.41	0.52	0.36	0.45	—	0.41	0.38	0.33	0.33	0.52
Less structured human ratings																				
16. Adaptability	0.02	0.12	−0.06	−0.08	0.01	0.45	0.30	0.23	0.37	0.21	0.61	0.30	0.30	0.35	0.31	—	0.44	0.45	0.41	0.40
17. Composure	0.10	0.09	0.09	0.05	0.03	0.29	0.46	0.33	0.41	0.29	0.41	0.61	0.44	0.38	0.50	0.40	—	0.40	0.40	0.43
18. Dependability	0.01	0.00	0.01	0.01	−0.16	0.35	0.32	0.48	0.29	0.25	0.32	0.38	0.57	0.38	0.38	0.24	0.32	—	0.35	0.37
19. Initiative	0.10	0.11	0.08	0.13	0.01	0.42	0.46	0.38	0.55	0.20	0.36	0.41	0.40	0.57	0.39	0.39	0.44	0.40	—	0.51
20. Problem solving	0.15	0.03	0.04	0.03	−0.03	0.39	0.35	0.36	0.34	0.29	0.30	0.48	0.35	0.37	0.50	0.34	0.44	0.19	0.41	—

Note: Orange highlight indicates convergent correlations/monotrait-heretomethod (MTHM); green highlight indicates heterotrait-monomethod (HTMM); white indicates heterotrait-heteromethod (HTHM).

condition, the AI interview scores correlated significantly more strongly with structured human scores than with less structured human scores for composure (Steiger's $z = 4.02$, one-tail $p < 0.01$) and problem solving (Steiger's $z = 2.41$, one-tail $p < 0.01$). However, no significant differences were observed for adaptability (Steiger's $z = -1.38$, one-tail $p = 0.08$), dependability (Steiger's $z = -0.38$, one-tail $p = 0.35$), or initiative (Steiger's $z = -0.17$, one-tail $p = 0.43$). In the honest condition, the AI interview scores correlated significantly more strongly with structured human scores than with less structured human scores for dependability ($z = 3.30$, one-tail $p < 0.01$) and problem solving ($z = 2.60$, one-tail $p < 0.01$) but not for adaptability ($z = -0.16$, one-tail $p = 0.44$), composure ($z = 1.15$, one-tail $p = 0.12$), or initiative ($z = -0.81$, one-tail $p = 0.21$). Overall, the AI scores tended to converge more strongly with the structured than less structured human interview scores, whether in honest or faking conditions, supporting Hypothesis 2.

Regarding Hypothesis 3 (i.e., that verbal ability correlates positively with interview score improvement), as reported in Table 5 (Online Supplement Table S1 reports equivalent results at the question/construct level, with similar findings), verbal ability correlated significantly with the RADS for overall self-report scores ($r = 0.26$, $p < 0.01$), including for all of the constituent self-report scales (rs ranging from 0.17 to 0.23, ps ≤ 0.05) except adaptability ($r = 0.16$, $p > 0.05$). However, verbal ability did not correlate significantly with the RADS for the overall AI interview scores ($r = 0.01$), the structured human interview scores ($r = 0.06$), or the less structured human interview scores ($r = 0.09$). Verbal ability only correlated significantly with question-level improvement for less structured human ratings of adaptability ($r = 0.18$, $p < 0.05$). Thus, verbal ability related to score improvement on self-reports but not AI, structured human, or less structured human interview scores—thus largely failing to support Hypothesis 3.

Addressing Hypothesis 4 (i.e., that ATIC correlates positively with interview score improvement), ATIC did not correlate significantly with the RADS for the AI interview scores ($r = 0.01$), the structured human interview scores ($r = 0.02$), or the less structured human interview scores ($r = 0.05$; Table 5). ATIC only correlated significantly with question-level improvement for less structured human ratings of problem solving ($r = 0.19$, $p < 0.01$). Thus, Hypothesis 4 was not supported.

For Hypothesis 5 (i.e., that honest impression management correlates positively with interview score improvement), Table 5 reports the correlation between the RADS for impression management tactics and the RADS for each scoring method.[2] The RADS for honest defensive impression management correlated positively with the AI interview score RADS ($r = 0.18$, $p < 0.05$) and the less structured human rating RADS ($r = 0.20$, $p < 0.01$) but not with the structured human rating RADS ($r = -0.04$, $p > 0.05$). Additionally, the RADS for deceptive extensive image creation correlated negatively with the AI interview score RADS ($r = -0.14$, $p < 0.05$). No other impression management behaviors correlated significantly with interview RADS. Thus, Hypothesis 5 was only supported for honest defensive tactics for AI and less structured human interview scores.

Regarding Research Question 1 (i.e., how responses changed across experimental conditions), we calculated the experimental condition's correlation with the LIWC variables and readability statistics (which returns p values equivalent to t-tests). The Online Supplement Table S2 reports the results, which we summarize briefly here because so few variables returned significant results. For LIWC, only two variables gave $p < 0.05$: participants used a lower proportion of filler words ($r = -0.12$) and anger words ($r = -0.14$) in the applicant-like condition, compared to the honest condition. The remaining 78 LIWC variables had $p > 0.05$. Similarly, neither Flesch's reading ease ($p = 0.14$) nor FORCAST ($p = 0.17$) exhibited significant differences between the two conditions. Thus, in line with the honest impression management results, we can tentatively suggest that interviewees strove to present themselves more professionally (i.e., fewer filler and anger words) in the applicant-like condition, but the effects were quite small.

5 | Discussion

Researchers and practitioners are increasingly interested in scoring asynchronous interviews with AI. However, we still know little about how they might be affected by deliberate attempts at impression management and faking. Our study suggests that AI scored interviews—when trained to replicate human ratings on BARS—function similarly to human ratings made on BARS when participants are instructed to respond like applicants. The score inflation on AI and human interview ratings were similar and tended to be considerably smaller than

TABLE 5 | *Correlations Between Verbal Ability, ATIC, Impression Management RADS, and Overall RADs.*

	Verbal ability	ATIC	Impression management tactics' RADS					
			HSP	HD	DSIC	DEIC	DIP	NV
Self-report RADS	0.26**	0.06	0.26**	0.09	−0.07	−0.05	−0.01	0.08
AI interview ratings RADS	−0.01	0.01	0.11†	0.18*	0.03	−0.14*	0.00	0.09
Structured human ratings RADS	0.06	0.02	0.07	−0.04	0.09	−0.09	−0.01	0.10
Less structured human ratings RADS	0.09	0.05	0.09	0.20**	0.07	−0.06	−0.02	0.12†

Note: RADS = regression-adjusted difference score. ATIC = ability to identify criteria. HSP = honest self-promotion. HD = honest defensive. DSIC = deceptive slight image creation. DEIC = deceptive extensive image creation. DIP = deceptive image protection. NV = nonverbal.
*$p < 0.05$; **$p < 0.01$.
†$p < 0.10$ (one-tailed). Verbal ability $N = 105$, ATIC $N = 104$, all other variables $N = 152$.

the score inflation observed on self-report measures. Additionally, the interview ratings' discriminant validity and method variance was not as negatively affected by the applicant-like responses as were self-reports. In contrast to prior research, we found that neither verbal ability nor ATIC related to interview score improvements when moving from the honest- to applicant-like condition. Honest defensive impression management tactics correlated with score increases in AI and less structured human ratings. And finally, we found very few differences in verbal behavior between the honest- and applicant-like conditions, although the differences that emerged do suggest more effective, honest impression management in the applicant-like condition.

5.1 | Theoretical and Practical Implications

The AI scored interviews exhibited shifts between the honest- and applicant-like conditions that were similar in magnitude to those observed for structured human ratings. Given that the AI was trained on structured human ratings, this is not surprising and suggests that this type of ML model is no more susceptible to socially desirable responding than structured human ratings. Meanwhile, the applicant-like condition substantially increased both the scores achieved on and the method variance of the self-reports. Together, these results further support the superiority of interviews over self-reports in the presence of applicant faking (e.g., Van Iddekinge et al. 2005) and extended those findings to include AI scored interviews.

While researchers have recently recognized that self-reported traits and skills often have close analogues (Soto et al. 2021), our results suggest that self-reports are not interchangeable with other measurement methods. Similar to how self-reports and other-ratings of personality can represent distinct constructs (McAbee and Connelly 2016), self-reports and interview ratings are empirically and conceptually distinct. Indeed, in both the honest- and applicant-like conditions, we found minimal convergence between self-reports and interview ratings (whether from AI or humans; maximum $r = 0.15$). Thus, self-reports and interview ratings both hold potential for providing valuable information about applicants that could contribute incremental validity beyond the other source of information.

Interviewee abilities have been shown to relate to both interview performance and improvement when moving from honest to applicant-like conditions (e.g., Buehl et al. 2019; Buehl and Melchers 2017; Ingold et al. 2015; König et al. 2007; Melchers et al. 2009). All prior studies of this phenomenon focused on face-to-face interviews. Consistent with prior work, we found that verbal ability and ATIC related to interview performance—both AI and human ratings. However, deviating from prior work, in our study neither ATIC nor verbal ability related to interview score improvement (although verbal ability related to self-report score improvement). Only two studies have related abilities to interview score *improvement*. Both were conducted by the same author teams, found relatively weak correlations ($r = 0.16$ for ATIC; $r = 0.19$ for cognitive ability, as measured with the Wonderlic), and had small sample sizes ($Ns \leq 111$) (Buehl et al. 2019; Buehl and Melchers 2017). In Buehl et al. (2019), the correlation between interview RADS and ATIC was only significant at $p < 0.05$ for a one-tailed test. Sampling error combined with publication bias may have resulted in false positive findings, or the nature of asynchronous interviews may differ from face-to-face interviews, such that the same effects do not occur. For example, asynchronous interviews are devoid of certain context cues (e.g., interviewer nonverbal cues) that could be present and serve as feedback in a synchronous interaction. Alternatively, we detail study limitations below that may have contributed to our failure to replicate the effect of interviewee abilities on interview score improvements.

Regarding impression management, only honest defensive tactics related to interview score improvement, and they did not relate to improvement on structured human ratings. Meanwhile, deceptive extensive image creation led to receiving lower AI interview scores in the applicant-like condition. These results align with prior research that found honest—but not deceptive—impression management tactics relate to interview performance (Ho et al. 2021; Peck and Levashina 2017). Notably, the nature of impression management differs in asynchronous interviews from face-to-face, given that other-focused tactics—such as ingratiation—are not relevant. There is a need for a greater understanding of the uses and influences of impression management tactics in asynchronous interviews, given their distinct, noninteractive nature.

Finally, we found that only the proportion of filler and anger-related words differed between the honest- and applicant-like conditions. Overall, we found very few differences in verbal behavior across the experimental conditions. This suggests that impression management tactics (which tended to increase when moving from the honest- to applicant-like condition) can be quite subtle and are not likely to be clearly detected from outside observers or by simple, behavioral measurements. This also holds implications for the literature on detecting faking and impression management, given that if behavior does not obviously differ between the two, then detecting socially desirable responding and faking will be difficult (if not impossible). Notably, we did not examine nonverbal and paraverbal behaviors given that the AI and structured human interview ratings focused exclusively on the content of interviewee responses.

5.2 | Limitations and Future Research

As with most studies on this topic, our study did not involve real job applicants. Thus, the behaviors exhibited in the honest- and applicant-like conditions may not exactly mirror the behaviors exhibited by real job applicants. Notably, the participants in the present study were required to complete both interview sessions (i.e., at Time 1 and Time 2) to receive payment for participation. In general, this resulted in high-effort participants and quality interview responses. Additionally, incentives were provided to encourage effortful responding in the applicant-like condition, but it would still benefit this area of research to gather more data from real job applicants. The observed effects of verbal ability and ATIC may have been attenuated in our study because they were measured approximately 1 year after the interviews were completed. For example, the participants' true levels of verbal ability and ATIC could

have changed over time, or this could increase measurement error due to differing occasions of measurement. Both variables still related to interview performance, but this attenuation may have prevented their relating to interview RADS. However, it may also be that asynchronous interviews function differently, such that these abilities do not have as substantive of an influence on score improvement as they do in face-to-face interviews. Future research is needed to tease apart these explanations.

In terms of study instructions, some researchers opted to instruct participants to "fake good" when examining faking (e.g., Zickar and Robie 1999), while others—like in our study—instructed participants to perform as an applicant (e.g., Buehl and Melchers 2017; Van Iddekinge et al. 2005). These two sets of instructions address different questions (Robie et al. 2001; Smith and Ellingson 2002). The "fake good" approach addresses the maximum extent that scores can be inflated in an assessment, while applicant-like instructions provide aim to elicit score inflation comparable to real applicants (Van Iddekinge et al. 2005). Our aim was to investigate how much AI-scored interviews are prone to realistic response inflation, but future work could also examine how extensively participants are able to inflate their scores when instructed explicitly to fake good.

Relatedly, participants in our study were not instructed that their interview would be scored by an AI system. Thus, their tactics for improving their score may differ from when they are aware that they are being evaluated by an AI scoring model. In particular, one can imagine applicants attempting to "game" such systems by providing responses that are nonsensical to humans but that contain keywords and buzzwords that they believe will lead to a high AI score. To date, no research has examined whether AI systems can be gamed with such adversarial responses. To the extent that they can, it raises concerns about the validity of AI scores in general. So, while future research is needed on this topic, organizations can also easily prevent such "gaming" of AI systems by ensuring that there is a "human in the loop" (e.g., Downes et al. 2023) who reviews each interview response—however briefly—to ensure that it is effortful. However, even a human in the loop may not be able to distinguish human-generated responses from responses generated by large language models (e.g., such as those that power ChatGPT), raising fresh concerns for the use of digitally-mediated interviews in high-stakes settings (Canagasuriam and Lukacik 2024).

On a related note, it may be possible to design algorithms that minimize score inflation in response to faking and IM. For example, it may be possible to use multi-objective optimization to minimize the influence of IM behaviors or to partial out IM variance from interview ratings. Future work could address this area of inquiry. Note, however, that some score inflation *should* occur—otherwise it suggests that the AI is not validly detecting behavioral differences across different interviews.

6 | Conclusion

Overall, AI interview ratings performed similarly to structured human ratings when comparing their psychometric properties in honest and applicant-like conditions. This suggests that AI interview ratings trained on structured human ratings are no more prone to applicant faking than are structured human ratings, further supporting their use in high-stakes settings. Unexpectedly, neither verbal ability nor ATIC related to interview score improvement in our study, suggesting that the score inflation may function differently in asynchronous interviews compared to face-to-face interviews. Honest defensive impression management tactics—but not deceptive ones—also led to increased interview scores in some cases. Future research could examine how inputs meant to game the AI system—whether generated by humans or large language models—perform when scored by an AI system.

Conflicts of Interest

Josh Liff and Colin Willis are employees of HireVue Inc., who sells access to algorithms for automatically scoring interviews. However, Louis Hickman collected the data and ran all analyses, except for the automatic interview scores which were provided by HireVue Inc., blind to the study conditions.

Data Availability Statement

The data that support the findings of this study are openly available on OSF at https://osf.io/dhwgm/.

Endnotes

[1]Note that this is not the case for large language models such as GPT and Llama-even when temperature is zero.

[2]Note that the raw impression management scores in the faking condition exhibit similar, but slightly attenuated, relationships with the RADS for each scoring method. We consider the IM RADS more meaningful, given that IM could already have been high in the honest condition, so the change in IM is more likely to explain the change in interview scores than the raw level of IM tactics.

References

Arthur, W., E. Hagen, and F. George. 2021. "The Lazy or Dishonest Respondent: Detection and Prevention." *Annual Review of Organizational Psychology and Organizational Behavior* 8: 105–137.

Auer, E. M. L. 2018. Detecting Deceptive Impression Management Behaviors in Interviews Using Natural Language Processing *[Master's thesis]*. Old Dominion University. https://doi.org/10.25777/yx69-dy97.

Barrick, M. R., J. A. Shaffer, and S. W. DeGrassi. 2009. "What You See May not be What You Get: Relationships Among Self-Presentation Tactics and Ratings of Interview and Job Performance." *Journal of Applied Psychology* 94, no. 6: 1394–1411.

Bobko, P., P. L. Roth, and C. Bobko. 2001. "Correcting the Effect Size of D for Range Restriction and Unreliability." *Organizational Research Methods* 4, no. 1: 46–61.

Borman, W. C., U. C. Kubisiak, and R. J. Schneider. 1999. "Work styles." In *An Occupational Information System for the 21st Century: The Development of O*NET*, Edited by N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, and E. A. Fleishman, 213–226. American Psychological Association. https://doi.org/10.1037/10313-012.

Bourdage, J. S., N. Roulin, and R. Tarraf. 2018. "'I (Might be) Just That Good': Honest and Deceptive Impression Management in Employment Interviews." *Personnel Psychology* 71, no. 4: 597–632.

Buehl, A. K., and K. G. Melchers. 2017. "Individual Difference Variables and the Occurrence and Effectiveness of Faking Behavior in Interviews." *Frontiers in Psychology* 8: 228017.

Buehl, A. K., K. G. Melchers, T. Macan, and J. Kühnel. 2019. "Tell me Sweet Little Lies: How Does Faking in Interviews Affect Interview Scores and Interview Validity?" *Journal of Business and Psychology* 34, no. 1: 107–124.

Burns, G. N., and N. D. Christiansen. 2011. "Self-Efficacy in the Workplace: Linking Personality to Domain-Specific Efficacy Beliefs." *International Journal of Selection and Assessment* 19, no. 4: 429–434.

Campion, M. C., M. A. Campion, E. D. Campion, and M. H. Reider. 2016. "Initial Investigation into Computer Scoring of Candidate Essays for Personnel Selection." *Journal of Applied Psychology* 101, no. 7: 958–975. https://doi.org/10.1037/apl0000108.

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: e12491. https://doi.org/10.1111/ijsa.12491.

Canagasuriam, D., and N. Roulin. 2021. "The Effect of Organizational Culture on Faking in the Job Interview." *Personnel Assessment and Decisions* 7, no. 1: 8.

Caylor, J. S., and T. G. Sticht. 1973. "Development of a Simple Readability Index for Job Reading Material." *Paper presented at the Annual Meeting of the American Educational Research Association* 1: 1. https://files.eric.ed.gov/fulltext/ED076707.pdf.

Downes, P. E., T. B. Harris, and D. G. Allen. 2023. "Getting From Valid to Useful: End User Modifiability and Human Capital Analytics Implementation in Selection." *Human Resource Management* 62, no. 6: 917–932.

Goldberg, L. R., J. A. Johnson, H. W. Eber, et al. 2006. "The International Personality Item Pool and the Future of Public-Domain Personality Measures." *Journal of Research in Personality* 40: 84–96.

Gordon, M. E., and R. H. Gross. 1978. "A Critique of Methods for Operationalizing the Concept of Fakeability." *Educational and Psychological Measurement* 38, no. 3: 771–782.

Griffin, B. 2014. "The Ability to Identify Criteria: Its Relationship With Social Understanding, Preparation, and Impression Management in Affecting Predictor Performance in a High-Stakes Selection Context." *Human Performance* 27, no. 2: 147–164.

Hickman, L., N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo. 2022. "Automated Video Interview Personality Assessments: Reliability, Validity, and Generalizability Investigations." *Journal of Applied Psychology* 107, no. 8: 1323–1351. https://doi.org/10.1037/apl0000695.

Hickman, L., L. Tay, and S. E. Woo. 2025. "Are Automated Video Interviews Smart Enough? Behavioral Modes, Reliability, Validity, and Bias of Machine Learning Cognitive Ability Assessments." *Journal of Applied Psychology* 110, no. 3: 314–335. https://doi.org/10.1037/apl0001236.

Higgins, C. A., and T. A. Judge. 2004. "The Effect of Applicant Influence Tactics on Recruiter Perceptions of Fit and Hiring Recommendations: A Field Study." *Journal of Applied Psychology* 89, no. 4: 622–632.

Ho, J. L., D. M. Powell, and D. J. Stanley. 2021. "The Relation Between Deceptive Impression Management and Employment Interview Ratings: A Meta-Analysis." *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement* 53, no. 2: 164–174. https://doi.org/10.1037/cbs0000223.

Hogan, J., R. Hogan, and T. Murtha. 1992. "Validation of a Personality Measure of Managerial Performance." *Journal of Business and Psychology* 7: 225–237.

Holtrop, D., J. K. Oostrom, W. R. J. van Breda, A. Koutsoumpis, and R. E. de Vries. 2022. "Exploring the Application of a Text-To-Personality Technique in Job Interviews." *European Journal of Work and Organizational Psychology* 31, no. 6: 799–816.

Huffcutt, A. I., S. S. Culbertson, and W. S. Weyhrauch. 2013. "Employment Interview Reliability: New Meta-Analytic Estimates by Structure and Format." *International Journal of Selection and Assessment* 21, no. 3: 264–276. https://doi.org/10.1111/ijsa.12036.

Hunter, J. E., and H. R. Hirsh. 1987. "Applications of meta-analysis." In *International Review of Industrial and Organizational Psychology*, edited by C. L. Cooper and I. T. Robertson, 321–357. Chichester, U.K.: Wiley.

Van Iddekinge, C. H., P. H. Raymark, and P. L. Roth. 2005. "Assessing Personality With a Structured Employment Interview: Construct-Related Validity and Susceptibility to Response Inflation." *Journal of Applied Psychology* 90, no. 3: 536–552.

Ingold, P. V., M. Kleinmann, C. J. König, K. G. Melchers, and C. H. Van Iddekinge. 2015. "Why do Situational Interviews Predict Job Performance? The Role of Interviewees' Ability to Identify Criteria." *Journal of Business and Psychology* 30: 387–398.

Jackson, D. N., S. V. Paunonen, and P. F. Tremblay. 2000. *Six Factor Personality Questionnaire: Technical Manual*. Port Huron, MI: Sigma Assessment Systems.

Jayaratne, M., and B. Jayatilleke. 2020. "Predicting Personality Using Answers to Open-Ended Interview Questions." *IEEE Access* 8: 115345–115355.

Kleinmann, M., P. V. Ingold, F. Lievens, A. Jansen, K. G. Melchers, and C. J. König. 2011. "A Different Look at Why Selection Procedures Work: The Role of Candidates' Ability to Identify Criteria." *Organizational Psychology Review* 1, no. 2: 128–146. https://doi.org/10.1177/2041386610387000.

König, C. J., K. G. Melchers, M. Kleinmann, G. M. Richter, and U. C. Klehe. 2007. "Candidates' Ability to Identify Criteria in Nontransparent Selection Procedures: Evidence From an Assessment Center and a Structured Interview." *International Journal of Selection and Assessment* 15, no. 3: 283–292.

Lang, J. W. B., and H. J. Kell. 2019. "General Mental Ability, Narrow Abilities, and the Efficacy of Specific Aptitudes in Employee Selection: Practical and Theoretical Issues." *Journal of Applied Psychology* 104, no. 10: 1235–1255.

Leary, M. R., and R. M. Kowalski. 1990. "Impression Management: A Literature Review and Two-Component Model." *Psychological Bulletin* 107, no. 1: 34–47. https://doi.org/10.1037/0033-2909.107.1.34.

Levashina, J., and M. A. Campion. 2007. "Measuring Faking in the Employment Interview: Development and Validation of an Interview Faking Behavior Scale." *Journal of Applied Psychology* 92, no. 6: 1638–1656.

Levashina, J., C. J. Hartwell, F. P. Morgeson, and M. A. Campion. 2014. "The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature." *Personnel Psychology* 67, no. 1: 241–293.

Liff, J., N. Mondragon, C. Gardner, C. J. Hartwell, and A. Bradshaw. 2024. "Psychometric Properties of Automated Video Interview Competency Assessments." *Journal of Applied Psychology* 109, no. 6: 921–948. https://doi.org/10.1037/apl0001173.

McAbee, S. T., and B. S. Connelly. 2016. "A Multi-Rater Framework for Studying Personality: The Trait-Reputation-Identity Model." *Psychological Review* 123, no. 5: 569–591. https://doi.org/10.1037/rev0000035.

Melchers, K. G., U.-C. Klehe, G. M. Richter, M. Kleinmann, C. J. König, and F. Lievens. 2009. "I Know What You Want to Know": The Impact of Interviewees' Ability to Identify Criteria on Interview Performance and Construct-Related Validity." *Human Performance* 22: 355–374. https://doi.org/10.1080/08959280903120295.

Melchers, K. G., N. Roulin, and A. K. Buehl. 2020. "A Review of Applicant Faking in Selection Interviews." *International Journal of Selection and Assessment* 28, no. 2: 123–142.

Morgeson, F. P., M. A. Campion, R. L. Dipboye, J. R. Hollenbeck, K. Murphy, and N. Schmitt. 2007. "Reconsidering the Use of Personality

Tests in Personnel Selection Contexts." *Personnel Psychology* 60: 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x.

Peck, J. A., and J. Levashina. 2017. "Impression Management and Interview and Job Performance Ratings: A Meta-Analysis of Research Design With Tactics In Mind." *Frontiers in Psychology* 8: 231201.

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

Putka, D. J., H. Le, R. A. McCloy, and T. Diaz. 2008. "Ill-Structured Measurement Designs in Organizational Research: Implications for Estimating Interrater Reliability." *Journal of Applied Psychology* 93, no. 5: 959–981. https://doi.org/10.1037/0021-9010.93.5.959.

Robie, C., M. J. Zickar, and M. J. Schmit. 2001. "Measurement Equivalence Between Applicant and Incumbent Groups: An IRT Analysis of Personality Scales." *Human Performance* 14, no. 2: 187–207. https://doi.org/10.1207/S15327043HUP1402_04.

Roch, S. G., D. J. Woehr, V. Mishra, and U. Kieszczynska. 2012. "Rater Training Revisited: An Updated Meta-Analytic Review of Frame-of-Reference Training." *Journal of Occupational and Organizational Psychology* 85, no. 2: 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x.

Rottman, C., C. Gardner, J. Liff, N. Mondragon, and L. Zuloaga. 2023. "New Strategies for Addressing the Diversity-validity Dilemma With Big Data." *Journal of Applied Psychology* 108, no. 9: 1425–1444. https://doi.org/10.1037/apl0001084.

Roulin, N., and D. M. Powell. 2018. "Identifying Applicant Faking in Job Interviews." *Journal of Personnel Psychology* 17, no. 3: 143–154. https://doi.org/10.1027/1866-5888/a000207.

Schneider, W. J., and D. A. Newman. 2015. "Intelligence is Multidimensional: Theoretical Review and Implications of Specific Cognitive Abilities." *Human Resource Management Review* 25, no. 1: 12–27.

Smith, D. B., and J. E. Ellingson. 2002. "Substance Versus Style: A New Look at Social Desirability in Motivating Contexts." *Journal of Applied Psychology* 87, no. 2: 211–219. https://doi.org/10.1037/0021-9010.87.2.211.

Soto, C. J., and O. P. John. 2017. "The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model With 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power." *Journal of Personality and Social Psychology* 113: 117–143.

Soto, C. J., C. M. Napolitano, and B. W. Roberts. 2021. "Taking Skills Seriously: Toward an Integrative Model and Agenda for Social, Emotional, and Behavioral Skills." *Current Directions in Psychological Science* 30, no. 1: 26–33.

Thompson, I., N. Koenig, D. L. Mracek, and S. Tonidandel. 2023. "Deep Learning in Employee Selection: Evaluation of Algorithms to Automate the Scoring of Open-Ended Assessments." *Journal of Business and Psychology* 38, no. 3: 509–527.

Tsai, W. C., C. C. Chen, and S. F. Chiu. 2005. "Exploring Boundaries of the Effects of Applicant Impression Management Tactics in Job Interviews." *Journal of Management* 31, no. 1: 108–125.

Vrij, A., P. A. Granhag, and S. Porter. 2010. "Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection." *Psychological Science in the Public Interest* 11, no. 3: 89–121.

Zickar, M. J., and C. Robie. 1999. "Modeling Faking Good on Personality Items: An Item-Level Analysis." *Journal of Applied Psychology* 84, no. 4: 551–563. https://doi.org/10.1037/0021-9010.84.4.551.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.