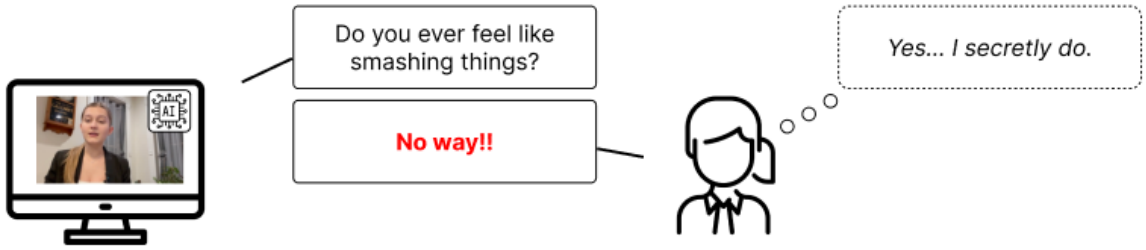
AI is Watching, Be Cool!

Maria DeCaro, Johnna Liu, and Stephen Dong



Original Experiment:

In 1989, Martin and Nagao [1] conducted a pioneering study to investigate how different interview formats influenced job applicants' responses, specifically focusing on social desirability and honesty. The experiment involved 103 undergraduate students who were asked to imagine applying for either a high-status management trainee role or a low-status clerk position. Four interview conditions were tested: computerized, paper-and-pencil, face-to-face with a "warm" interviewer, and face-to-face with a "cold" interviewer.

To measure social desirability, they used the Marlowe-Crowne Social Desirability Scale—a psychological assessment consisting of 33 true/false questions designed to gauge the tendency of individuals to present themselves in a favorable light by responding in socially approved ways. They assessed honesty by comparing self-reported SAT and GPA scores between the different groups. The findings revealed that non-social formats, such as computerized or paper-and-pencil interviews, elicited lower social desirability scores and more accurate self-reports. However, these formats also led to greater applicant resentment, particularly for high-status positions. The study underscored a trade-off between eliciting honest responses and maintaining positive applicant perceptions.

Replicated Experiment:

Our team aimed to expand on the original study by examining how modern AI influences applicant responses in professional interviews. To replicate and build upon the initial experiment, we selected one social and one non-social interview format to provide a foundation for comparison, exploring where AI fits within this spectrum. Specifically, we included the original

study's face-to-face interview with a "warm" human interviewer and a computerized interview format, while introducing a new AI-led condition to reflect technological advancements.

Participants began by completing a pre-interview survey that collected demographic information, including job interview experience, work history, and technological familiarity. After the survey, participants were randomly assigned to one of three interview conditions: a human-led, face-to-face "warm" interview, a neutral computerized interview, or a human-like AI-led interview. As an incentive to do their best in the interview, we informed subjects that the best three performing interviewees (and thus most likely to be hired) would receive a $15 gift card, similar to the original experiment where the participants were told the top 20% of those interviewed would receive double extra credit points (for a course).

For the "warm" interviewer condition, we closely followed the original study's approach. The interviewer (Maria) maintained a positive demeanor by smiling and using non-verbal cues, such as nodding, to encourage participants. For the computerized interview, we adapted the original text-based format to a video format, reflecting current practices and ensuring participants in all conditions responded verbally to interview questions. Standardizing verbal responses across all conditions reduced the potential for bias introduced by the mode of response. In the computerized format, participants viewed a black screen while questions were read aloud, further minimizing the influence of interviewer bias.

The AI-led interview condition was implemented using the HeyGen platform to create realistic AI-generated videos. These videos featured Maria's face and voice to align with the face-to-face interview while maintaining consistency across conditions. Participants were explicitly informed that the videos were AI-generated. This condition was designed to reflect the increasing role of AI in recruitment, which may shape applicant responses through perceptions of impartiality, efficiency, and emotional detachment.

Treatment Conditions

1. Human-interviewer 2. Computer-Interviewer 3. AI-Interviewer

After completing their assigned interview, participants were asked to complete a post-interview survey, similar to the original experiment, to assess comfort, perceived performance, and resentment toward the interview method. Social desirability was measured using the Marlowe-Crowne Social Desirability Scale, while honesty was assessed by comparing self-reported GPAs between the groups. We hypothesized that AI-led interviews would balance honesty with positive applicant perceptions, presenting a viable alternative to traditional human and computerized formats.

Sampling Technique:

Participants Demographic Table

Gender	Number of Participants	Average Work Experience	Average Job Interview Experiences	Technology Background
Male	18	1.889	14.444	88.9%
Female	12	0.708	8.333	58.3%
Total	30	1.417	12.000	76.7%

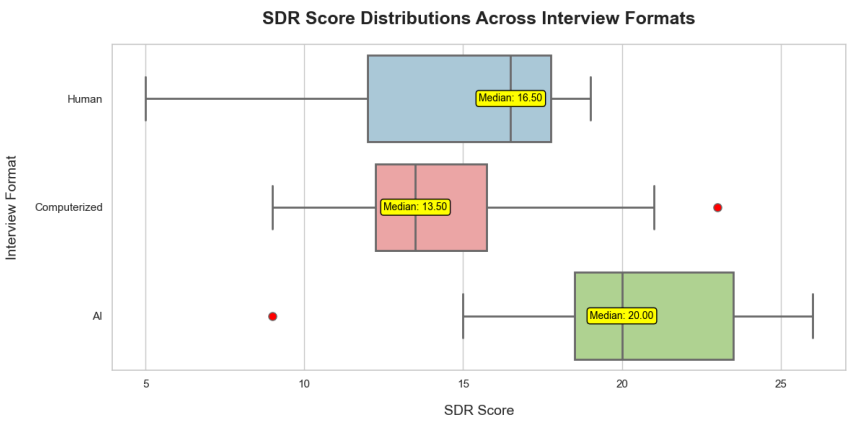
We recruited 30 participants through a combination of friend networks and convenience sampling, targeting Cornell Tech students as our population of interest. This sampling method allowed us to quickly gather the target sample size. In the table above, we highlight the participant breakdown from our experiment, extracted from the pre-interview survey. Our sample included 18 male and 12 female participants. These participants had an average work experience of 1.417 years and conducted an average of 12 interviews. 76% of the population identified as having a “technological background,” which we classified as majors including computer science, math and electronic engineering, which typically involve a stronger focus on coding and machine learning.

Results:

Mean and Variance for SDR and GPA

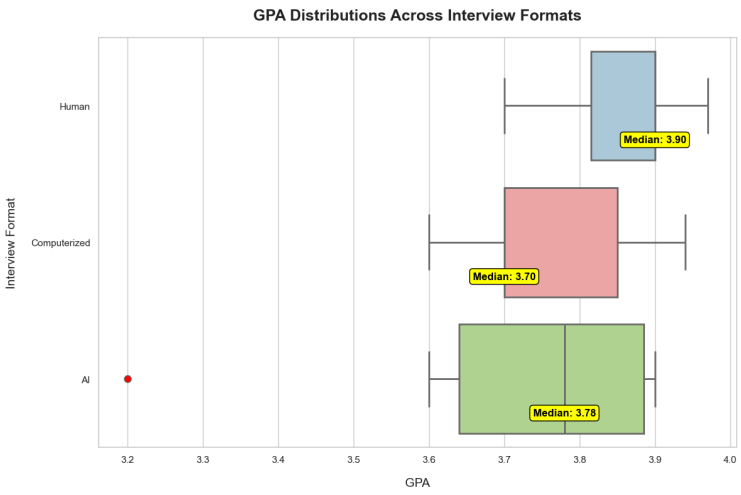
Interviewer	Median (SDR)	IQR (SDR)	Median (GPA)	IQR (GPA)
AI	20.00	5.00	3.78	0.25
Computer	13.50	3.50	3.70	0.15
Human	16.50	5.75	3.90	0.09
Combined	17.00	7.75	3.80	0.20
Max Scores	33.00		4.00	

Our experiment yielded mixed results. In the above figure, we report the median and interquartile ranges of the participants' SDR scores and GPAs across each treatment condition. To capture the variability of the small dataset, we employed median metrics and used statistical tests to validate our observations. Participants reported GPAs aligned with our expectations: those in the AI-interviewer group reported median GPAs that fell between those of human-interviewer groups and computerized-interviewer groups. For the SDR scores, while we did find that human-interviewer groups had higher SDR than computerized-interviewer groups, we were surprised to find that the AI-interviewer group had the highest SDR scores.

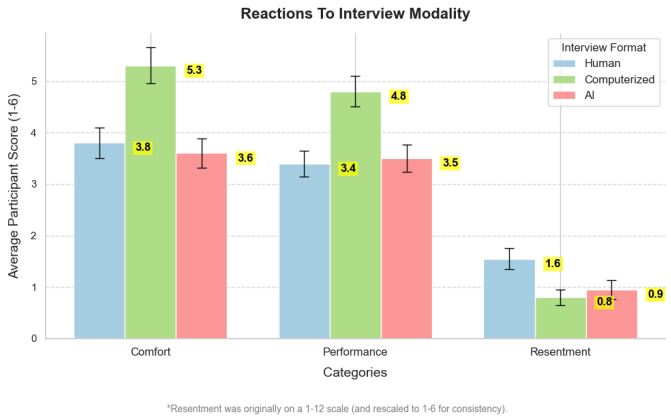


As we examine the boxplot of SDR scores, we find that the median for the AI group exceeds the maximum for the human group, and approaches the maximum for the computerized control group. While we can only see trends with this current distribution, what is interesting is that while a two-sided Mann-Whitney U test at a .05 significance level indicated no statistically significant difference between the human and computerized groups (p 0.8406), it indicated a statistically significant difference between the computerized and AI interview groups (p=0.0371) and the human and AI interview groups (p = 0.0122).

On the other hand, we saw that the GPA distributions were much closer. We were able to replicate the GPA being reported higher in the human-led group, and lower in computerized (non-social interview groups). In addition, as we expected, we found the self-reported GPA to fall in between these two original conditions when an AI interviewer was in play. When we applied the two sided Mann-Whitney U test we were not able to find any statistical significance at a .05 significance level (human vs. computerized: 0.0804, human vs AI: 0.071, computerized vs AI: .939). The relatively low p-values across the human group might indicate that more data might reveal a different story.



Finally, within the post-interview experiment form, we asked participants to rank on a Likert scale how comfortable they felt during the interview, how they perceived their interview performance, and whether they felt any resentment about the interview. Contrary to the original experiment, which found that participants bore more resentment towards computerized interviews, we found that participants bore the least resentment to computerized interviews, the most with human-led interviews, and in-between with AI-led interviews. Across the other metrics, we found that participants felt most comfortable and performed the best across the computerized interviews, and approximately equally less so across the human and AI-led interviews.



Conclusion and Discussion:

Our replication study extends Martin and Nagao's seminal work into the AI era, revealing both validating insights and surprising departures from their original findings. While their study showed that non-social interview formats consistently reduced social desirability responses, our results paint a more complex picture in today's technological landscape.

Besides successfully replicating the results of original study that human-led interviews resulted in higher SDR scores and inflated GPAs compared to computerized interviews, the most striking finding was that AI-led interviews produced unexpectedly high socially desirable responding (SDR) scores (median 20.0) compared to both human-led (16.5) and computerized (13.5) interviews. This suggests that AI interviews may introduce novel psychological dynamics that intensify rather than reduce social desirability bias, potentially due to AI's perceived authority or uncertainty about AI evaluation methods. However, when examining verifiable information like GPA reporting, AI interviews yielded results (median 3.78) that fell between human-led (3.90) and computerized formats (3.70), indicating that AI might occupy a unique middle ground in encouraging truthful self-disclosure.

Our findings regarding interview reactions also departed from the original study. While Martin and Nagao found higher resentment toward computerized formats for high-status positions, our participants showed greater resentment toward human-led interviews (3.1) compared to both AI (1.9) and computerized formats (1.6). This shift likely reflects evolving attitudes toward technology in professional settings over the past decades.

Several limitations of our study warrant consideration. First, our convenience sampling of Cornell Tech students may have skewed results due to their higher technological literacy and familiarity with AI systems. These participants may have different attitudes toward AI interviews compared to the general population or older professionals. Second, the use of HeyGen for generating AI interviews may have introduced platform-specific effects that might not generalize to other AI interview systems. Third, our relatively small sample size (30 participants) and time constraints may have limited the statistical power of our findings.

Additionally, the study's focus on a specific demographic (primarily young, tech-savvy graduate students) raises questions about the generalizability of our findings across different age groups, industries, and cultural contexts. Future research should examine these effects across more diverse populations and professional settings.

These findings have important implications for both research and practice. For researchers, the complex relationship between AI interviews and social desirability behavior suggests the need for deeper investigation into the psychological mechanisms at play, particularly regarding how different AI personalities or interaction styles might affect candidate responses. The disconnect between SDR scores and actual honesty metrics also warrants further exploration. For organizations, our results suggest that while AI interviews may offer certain advantages in terms of candidate comfort and standardization, they require careful implementation to manage their tendency to increase social desirability bias.

Our study ultimately suggests that while AI interviews present certain advantages, their implementation requires careful consideration of multiple factors to ensure effectiveness. As organizations increasingly adopt AI interview systems, understanding these nuanced effects

becomes crucial for developing interview processes that balance efficiency with fairness and accuracy. The evolution of interview dynamics since Martin and Nagao's original work highlights the need for continued research as technology advances and reshape our understanding of human-AI interaction in professional contexts.

Original Study:

Christopher L. Martin and Dennis H. Nagao. 1989. Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology* 74, 1: 72–80. https://doi.org/10.1037/0021-9010.74.1.72

Appendix:

AI Interview:
https://drive.google.com/file/d/1hLXZtza8b0H6mxVFIIpr9rsGze2zyf4n/view?usp=drivesdk

Computerized Interview:
https://drive.google.com/file/d/1PV3OzTlrrfRUWixMqtIZe-CrSl2eKDt5/view?usp=sharing

Marlowe Crowne Scale:
https://drive.google.com/file/d/1HN5Kebca-eQMtvNRPfXuq86Y5aDMa2oF/view?usp=drivesdk

Pre-Interview Questionnaire:
https://docs.google.com/forms/d/e/1FAIpQLSe2aC-N7h46kSERM-lhk9j0oTGJjO0q17hqkpbkxqGE-my7Sw/viewform?usp=sf_link

Post-Interview Questionnaire:
https://docs.google.com/forms/d/e/1FAIpQLSdBV05509YeIboNnlSt-n27exCSkQTgcWXvwlDtQnZ8nRoENA/viewform?usp=sf_link

Attribution:

To design the graphics in this paper, we used the AI icon by *shin_icons*, monitor icon by *xnimrodx*, and the woman icon from *Vitaly Gorbachev*, all of which we retrieved from flaticons.com.