https://doi.org/10.35662/unine-thesis-3173

Université de Neuchâtel

Applicants' storytelling in behavioral interviews: Examining the roles of technology, culture and response framing in interview processes, outcomes, and criterion validity

PhD thesis submitted to the Faculty of Science
Institute of Work and Organisational Psychology
University of Neuchâtel
Switzerland

by

Elisabeth Germanier

Dissertation committee :

Prof. Adrian Bangerter, thesis director, University of Neuchâtel, Swirtzerland
Prof. Maike Debus, internal examiner, University of Neuchâtel, Switzerland
Prof. Markus Langer, external examiner, University of Freiburg, Germany

Defense: 3rd February, 2025

University of Neuchâtel, 2025

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame Elisabeth GERMANIER

Titre :

"Applicants' storytelling in behavioral interviews: Examining the roles of technology, culture and response framing on interview processes, outcomes, and criterion validity"

sur le rapport des membres du jury composé comme suit :

- **Prof. Adrian Bangerter,** directeur de thèse, Université de Neuchâtel, Suisse
- **Prof. Maike Debus,** Université de Neuchâtel, Suisse
- **Prof. Markus Langer,** University of Freiburg, Allemagne

Neuchâtel, le 6 mars 2025

Le Doyen, Prof. P. Brunner

Abstract

Job interviews are a way for recruiters to assess applicants' characteristics in a social encounter where recruiters and applicants meet and exchange information about each other. Yet, the emergence of technology-mediated interview formats such as asynchronous video interviews (AVIs) and artificial intelligence (AI) have transformed job interviews and pose new challenges and opportunities. Moreover, social behaviors are influenced by culture, thus governing one's behavior in job interviews. This thesis focuses on interview processes such as applicants' responses and outcomes, like recruiters' perceptions and evaluations, examining the role of response framing, technological innovations and cultural dynamics; it also investigates the ability of AVIs to predict job performance and the potential of AI in helping applicants in this process.

This thesis presents four studies that address these questions. Study 1 explored the impact of emotional framing in responses on recruiters' evaluations. Study 2 examined how interview media and individuals' culture affect applicants' responses and recruiters' perceptions of applicants. Building on this, Study 3 further examined the effects of AVIs versus face-to-face interviews on applicants' responses and performance and established how well AVI helps predict job performance. Finally, Study 4 developed AI models to analyze the content of applicants' responses, paving the way for automated coaching platforms for applicants.

Findings indicate that the emotional framing of responses plays a role in how recruiters perceive them. Also, both the interview medium and applicants' cultural background influence the content of responses and recruiters' evaluations. However, the response content and interview performance were mostly comparable across interview media. Results also indicate that AVIs can predict applicants' job performance as effectively as face-to-face interviews. Finally, our AI model demonstrates that it is possible to automate the analysis of response content from audio recordings.

This thesis helps to understand what affects applicants' responses and affects interview outcomes and encourages to further investigate the potential of AI to provide personalized interview training. It offers implications for interview research and practice and discuss its limits.

Keywords: job interviews, culture, asynchronous video interviews, storytelling responses, recruiters' evaluation, interview validity, AI-powered training.

Résumé

Les entretiens d'embauche permettent aux recruteur-euse-s d'évaluer les caractéristiques des candidat-e-s dans le cadre d'une situation sociale où tou-te-s se rencontrent et échangent des informations les un-e-s sur les autres. Cependant, l'émergence de formats d'entretien digitaux comme les entretiens vidéo asynchrones (*asynchronous video interview*, AVIs) et de l'intelligence artificielle (IA) a transformé les entretiens, posant de nouveaux défis et de nouvelles opportunités. De plus, les comportements sociaux sont influencés par les normes culturelles, déterminant ainsi les comportements de chacun-e lors d'entretiens d'embauche. Cette thèse se concentre sur les processus d'entretien comme les réponses des candidat-e-s et les résultats comme les perceptions et évaluations des recruteur-euse-s en examinant le rôle du cadrage des réponses, des innovations technologiques et des dynamiques culturelles ; elle examine également la capacité des AVIs à prédire la performance au travail et le potentiel de l'IA pour aider les candidats dans ce processus.

Quatre études ont été menées pour répondre à ces questions. L'étude 1 a exploré l'impact du cadrage émotionnel des réponses sur les évaluations des recruteur-euse-s. L'étude 2 a examiné la manière dont le média d'entretien et la culture de chacun-e affectent les réponses des candidat-e-s et la manière dont ils vont être évalués par les recruteur-euse-s. L'étude 3 a investigué les effets des AVIs par rapport aux entretiens face à face sur les réponses et la performance des candidat-e-s ainsi que leur capacité à prédire la performance au travail. Finalement, l'étude 4 a développé des modèles utilisant l'IA pour identifier le contenu des réponses directement depuis des enregistrements audios de réponses dans le but d'ouvrir la voie à des plateformes de coaching automatisées pour les candidat-e-s.

Les résultats ont montré que le cadrage émotionnel des réponses joue un rôle dans la manière dont les recruteur-euse-s les perçoivent et les évaluent. De plus, le médium d'entretien ainsi que les normes culturelles des candidat-e-s influencent le contenu des réponses et les perceptions des recruteurs. Néanmoins, les réponses des candidat-e-s et les évaluations des compétences sont comparables d'un médium à l'autre. De plus, les résultats montrent que les AVIs prédisent la performance au travail aussi bien que les entretiens face à face. Finalement, nos modèles utilisant l'IA démontrent qu'il est possible d'automatiser l'analyse du contenu des réponses à partir d'enregistrements audios de réponses.

Cette thèse permet de mieux comprendre de ce qui influencent les réponses des candidat-e-s ainsi que les résultats des entretiens d'embauche et encourage à investiguer le potentiel de l'IA pour la préparation personalisée des entretiens. Elle apporte des implications pratiques pour la recherche et la pratique tout en discutant de ses limites.

Mots clés: Entretien d'embauche, entretien vidéo asynchrones, réponses narratives, évaluations des recruteurs, validité des entretiens, entrainement assisté par l'IA

Acknowledgement

This thesis results from a journey made possible by the support of many remarkable individuals. Their encouragement has fueled my curiosity and helped me to explore new ideas. I want to take the opportunity to thank them.

I would first like to express my gratitude to my supervisor, Prof. Adrian Bangerter. Thank you for your constant support and for always giving me guidance that allowed me to explore my curiosity and creativity and turn it into something meaningful. Throughout my PhD, you have always been incredibly available, kind, and of good advice. I have truly appreciated it, and I feel lucky and honored to have worked with you. Thank you for guiding me through this journey!

I would also like to express my heartfelt thanks to a true gem, my colleague and friend, Koralie Orji. Your unconditional love and support, honesty, and wisdom have helped me grow and blossom in ways I hadn't thought before. Thank you for sharing this journey with me, with countless laughs and deep conversations about life and research.

I am grateful to everyone involved in this FNS project, starting with the project team: Marianne Schmid Mast, Laetitia Renier, Philip Garner, and Mutian He. Your diverse perspectives and expertise have been enriching. I owe special thanks to those who contributed to data collection, transcription, and coding for the studies in this thesis: Imen Kitar, Leïla Ouhamma, Fatlume Rashiti, Luisa Ferreira Araujo, Orlane Rota, Cécile Aubry, Raksha Gopalakrishnan, Célia Froidevaux, Théo Schöpfer, Koralie Orji, Sarah Pereira Dias, Pedro Semedo Mendes, Julie Kempter, Estelle Carrard, Valérie De Luca, Erika Ugolini, Salome Wimmer, Aricia Andrey, Alessia Garzilli, Mickael Amman, Walid Ezzine Charraf and Valeriia Kovaleva. I will forever cherish memories of funny meetings and the endless data collection.

I also thank the members of the Institute of Work and Organizational Psychology for their constructive feedback and refreshing lunch discussions. I would especially like to thank Adrian's past and present team, and to express my affection and gratitude to Garance Deschenaux. You are magic! Thanks to all, I feel fortunate to have had such a stimulating, caring, and constructive work environment.

A special thanks goes to my jury members, Markus Langer and Maike Debus, who found the time to read this thesis and come to Neuchâtel for the defense.

Lastly, I want to thank my friends Maël Theubet and Roald Brunell, my family, Thibault, Aska, and her two sisters, for their affection, support, and tolerance. I could experience every emotional season of this adventure unfiltered, from sharing hours of thrilling discussions about statistics to confessing my insecurities.

Contents

Introduction

> A crucial element in the maintenance of high productivity in both government and private industry is the selection of people with high ability for their jobs.
>
> **Hunter and Hunter (1984, p. 72)**

Choosing the right people who align with an organization's mission and values is crucial for its success. Good hires can positively impact both short-term and long-term outcomes. Conversely, poor hiring decisions can result in serious issues, such as high turnover, plummeting productivity, and poor morale, which are just the tip of the iceberg (Glassdoor Team, 2015; Hunter & Hunter, 1984; UNIKO Media Group, 2024). The cost of such selection mistakes can lead to significant financial losses that may wreck an organization. Given the adverse effects of poor hiring decisions, recruiters must employ effective and reliable methods to thrive in the talent acquisition landscape, ensuring they select applicants who can contribute positively to their vision and objectives. As such, job interviews stand out as the most widely used method to predict applicants' future job performance (Levashina et al., 2014; Posthuma et al., 2002).

In recent years, technological innovations have introduced new interview formats, like asynchronous video interviews (AVIs), and artificial intelligence (AI)-assisted evaluations. In AVIs, applicants engage in one-way interviews where they respond to pre-set questions online and submit their recorded responses to recruiters, which are evaluated later (Lukacik et al., 2022). Nevertheless, applicants often react negatively toward AVIs, *a fortiori* when implemented with AI, and think they have less chance to perform well in such settings (Basch et al., 2020; Langer et al., 2019, 2020). At the same time, only a few studies investigated how AI may help train applicants for future interviews (e.g., Hoque et al., 2013; Langer et al., 2016; Gebhard et al., 2019; Bangerter et al., 2023).

These innovations pose new challenges to our understanding of the dynamics of job interviews. One such challenge lies in the social nature of job interviews. Beyond being an instrument to assess applicants' characteristics, they also constitute social interactions where applicants and recruiters meet and influence one another impression through verbal and non-verbal behaviors (Bolino et al., 2008; Schlenker, 1980). The lack of real-time interaction in AVIs really shifts how applicants experience this influence process. Unlike face-to-face interviews, which offer a rich and interactive setting for impression management and rapport building, they lack recruiters' presence in AVIs. Furthermore, they are

deprived of naturally occurring feedback cues from recruiters. In this way, applicants are forced to self-monitor their responses, potentially leading to different impressions created on recruiters and interview performance.

A second challenge arising from the social aspect of interviews is culture. Culture plays a central role in setting social norms, ultimately defining behaviors considered appropriate in given social contexts (Liu et al., 2022). As such, one behaves in job interviews in accordance with one's cultural norms. How the self-image is managed in interviews can thus vary from culture to culture (Arseneault & Roulin, 2023). Moreover, AVIs offer a different type of social interaction that is likely to affect people differently depending on their culture (Griswold et al., 2022).

Facing these challenges, the main objective of my thesis is to deepen the understanding of how technological innovations, specifically AVIs and AI, and cultural differences affect the job interview processes and outcomes. To do so, I begin this thesis with a review of traditional job interviews. Specifically, I outline its psychometric foundations, social aspects, and the role of cultural differences (Chapter 1). Building on this review, I explore how applicants respond to questions in traditional face-to-face interviews, focusing on storytelling responses as they are optimal for creating lasting impressions (Chapter 2). In connection, I present Study 1, which explores the effect of applicants' response framing on interview outcomes. After providing a deepened understanding of face-to-face job interviews, I explore the technological changes (namely AVIs and AI) in job interviews in Chapter 3. I present Study 2 and Study 3, which investigate the effect of AVI and culture on various interview processes and outcomes. In the second part of that chapter, I explore how AI models can help train applicants to respond to interview questions. This leads to Study 4, which explores how AI can be used to analyze applicants' responses, particularly storytelling. Finally, this thesis concludes with a discussion of the findings of the presented studies, at the same time outlining their limitations and the opportunities for future research.

1 Job Interviews in Personnel Selection

> The selection process is a two-way interaction where applicants and organizations gather information about one another and react to this information while making employment decisions.
>
> **Bauer et al. (1998, p. 892)**

Job interviews have two main goals. First, they allow recruiters to evaluate further the fit of applicants who stood out from their resumes. Second, they offer a social encounter to initiate rapport building. In this context, recruiters try to build a positive image of their organization. Applicants, on the other hand, want to make a good impression to increase their chances of being hired for the job. Hence, job interviews are social interactions that revolve around the question-answer dynamic, where recruiters and applicants exchange information to assess fit from both sides (Bauer et al., 1998).

In this light, job interviews encompass two perspectives. First, from the psychometric perspective, they help assess applicants' characteristics, such as personality and competency, to predict applicants' future performance. Second, from a social perspective, they are social interactions where both parties attempt to influence each other's perceptions through positive self-presentation. They therefore engage in verbal and non-verbal impression management strategies. However, when it comes to social interactions, it is essential to take cultural dynamics into account, as one's behaviors are ruled by cultural norms (Liu et al., 2022). Hence, it is essential to identify these differences to ensure that interview practices remain fair and valid in multicultural contexts.

Because job interviews revolve around this dual psychometric and social nature, I develop these two perspectives in the next subchapters to show how they underpin the processes and outcomes of job interviews. This will help to understand later the role of response framing, technological innovation and cultural differences in the context of job interviews. To do so, I first present the psychometric perspective of job interviews, focusing on how behavioral interviews constitute a best practice (1.1). I then examine the social aspect of interviews, highlighting how self-presentation and impression management impact interview processes and outcomes (1.2). Finally, I explain the importance of considering cultural differences in job interviews, illustrating how cultural factors can affect applicants' and recruiters' behaviors (1.3).

1.1 Psychometric Foundations of Job Interviews

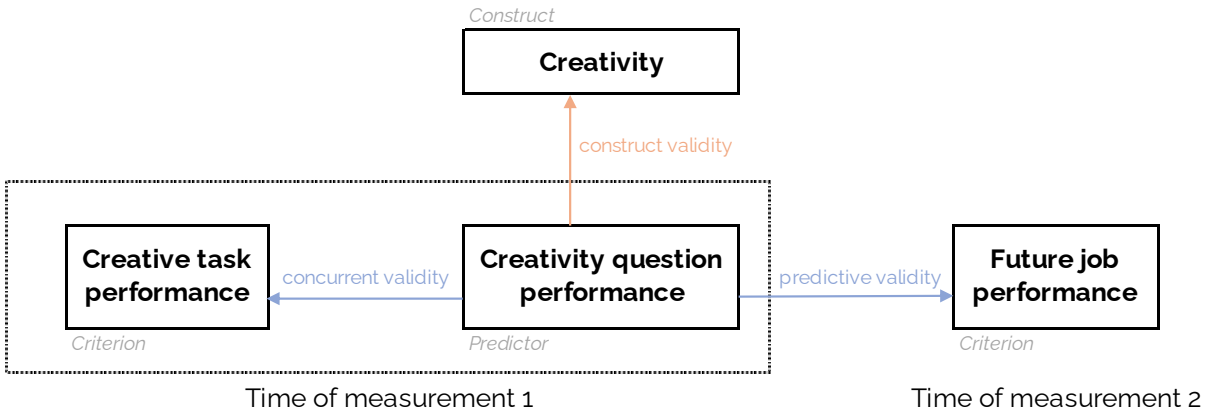
Job interviews can be approached in different ways. In the psychometric approach, the interview is a scientific process based on observable criteria and analytical reasoning (Roulin et al., 2012). In other words, the interview is an instrument aiming to select the best applicant for the job. It began to gain importance in the mid-20th century (Arvey & Campion, 1982; Schmidt & Hunter, 1998) in response to the limitations of the intuitive approach, which lacked robust and reliable selection methods supporting hiring decisions. The latter excessively relies on recruiters' subjective impressions and feelings, allowing various biases into evaluations (e.g., first-impression bias). The intuitive approach may thus sometimes lead to inaccurate decisions, and even when successful, it remains challenging to identify what led to that successful selection (Roulin et al., 2012). In contrast, the systematic approach advocated by the psychometric perspective seeks to quantify applicants' characteristics relevant to their performance in the job position and develop measures that can be compared across applicants. This approach facilitates knowledge sharing and helps future recruiters replicate success effectively.

In psychometry, a good instrument must be reliable and valid. Reliability looks at the ability of an instrument to produce consistent results when assessing the same characteristic in similar contexts (Roulin et al., 2012). For instance, a person that retakes the same test under similar conditions should get a comparable score. Validity looks at whether an instrument really measures what it is intended for. This can be done through construct validity or criterion validity (Cascio & Aguinis, 2014; Roulin et al., 2012). Construct validity refers to the extent to which an instrument accurately measures the targeted concept, known as the construct. Criterion validity evaluates how well the score of a test correspond to the outcomes or behaviors it is designed to predict, which is called the criterion. Criterion validity is further divided into two types: concurrent validity, where the criterion and test are measured simultaneously, and predictive validity, where the test forecasts a criterion that will occur in the future.

To illustrate interview different validities, let us consider an organization that interviews applicants for an art director position, where creativity is a key requirement. Recruiters ask questions to assess the applicants' creativity. After the interviews, the applicant with the highest creativity score is selected, assuming that this applicant will outperform others in the role. In this example, construct validity checks if applicants' responses to creativity questions truly reflect the construct of creativity. A high validity indicates that those responses provided an accurate picture of the applicant's creative abilities. As for the criterion-related validities, predictive validity is about figuring out

whether doing well on these questions predicts future success in the job. If applicants were asked to perform a creative task, such as pitching a concept for a brand with minimal preparation, the concurrent validity would assess the relationship between applicants' performance on this creative task and performance on the creativity interview questions (see Figure 1).

Figure 1: Criterion (Concurrent and Predictive) Validity and Construct Validity



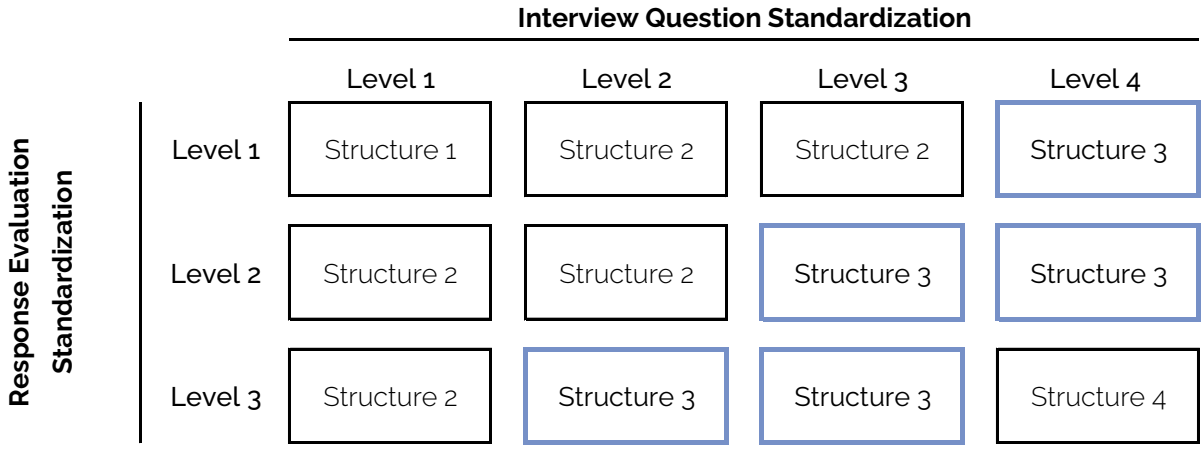
1.1.1 Job Interview Structure and Validity

Research on the psychometric properties of job interviews evidenced that interview structure is a crucial element for job interview reliability and predictive validity (e.g., Campion et al., 1997; Conway et al., 1995; Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Indeed, structured interviews have emerged as the most effective method for predicting future job performance, particularly compared to unstructured interviews (Sackett et al., 2022). Interview structure is defined as "any enhancement of the interview that is intended to increase psychometric properties by increasing standardization or assisting the interviewer in determining what questions to ask or how to evaluate responses" (Campion et al., 1997, p. 656). Structure is thus a multidimensional variable involving many possibilities to standardize the interview process. Campion et al. (1997) identified 15 ways to standardize both content and evaluations, among which the most frequently used by practitioners are developing questions based on job analysis, asking the same questions to applicants, asking better questions, using rating scales, rating each question, and training recruiters (Levashina et al., 2014).

Huffcutt and Arthur (1994) introduced a two-dimensional model that looks at structure as a continuous variable influenced by two main factors: the questions asked during

interviews and how those responses are evaluated. They identified four progressive levels of standardization for interview questions: 1) unstructured, which means there are no constraints; 2) limited constraints, where specific topics need to be covered; 3) pre-specified questions, which are set questions that allow for probing; and 4) full standardization, where all applicants answer the same questions in the same order without any probing. For response evaluations, there are three progressive levels: 1) a single overall evaluation of the entire interview, 2) multiple evaluations on pre-established criteria, and 3) an evaluation of each response with pre-established scales. Therefore, each interview integrates varying levels of standardization for both questions and evaluations, resulting in four combinations for structured interviews (see Figure 2).

Figure 2: The Four Levels of Interview Structure Adapted from Huffcutt & Arthur (1994)



Note. Reproduced with permission from American Psychological Association. No further reproduction or distribution is permitted. Light blue border indicates the ceiling effect of predictive validity at Structure 3.

Viewing structure as a multidimensional continuum, Huffcutt and Arthur (1994) found that the predictive validity increases with greater interview structuring. However, further structuring yields no substantial improvement to predictive validity beyond a certain point (Structure 3). This ceiling effect is important to consider when developing interviews, especially given the challenges of highly structured formats. Not only are they costly to develop, but both recruiters and applicants also express concerns about them. Indeed, recruiters are less favorable toward high structure because they perceive less flexibility and control, which is lower discretion and rapport-building (Lievens & De Paepe, 2004; Van der Zee et al., 2002). As for applicants, they perceive structured interviews as fairer but prefer the greater freedom in responding in unstructured interviews (Conway & Peneno, 1999; Kohn & Dipboye, 1998).

1.1.2 Behavioral Interviews and Past-Behavior Questions

Behavioral interviews are one specific type of structured interview, thus adopting a systematic approach. Such interviews are based on a job analysis collecting critical incidents — job situations where success critically relies on the employee's appropriate behaviors (Campion et al., 1997; Roulin et al., 2012). For each critical incident, recruiters identify behaviors related to success and those related to failure and then develop interview questions to invite applicants to explain their behaviors in similar situations. Based on the behaviors described in responses, recruiters assess applicants' mastery of the competencies of interest (Roulin et al., 2012). According to a survey, 75% of recruiters assess applicants' characteristics using behavioral questions (McLaren, 2019).

There are two types of behavioral questions based on the analysis of critical incidents: situational (e.g., "Imagine you are faced with a situation where you need to manage an angry client. How would you handle this situation?") and past-behavior questions (e.g., "Can you describe a situation where you had to manage an angry client?"; Roulin et al., 2012). Both question types ask applicants to describe their behaviors in a given situation but differ in the temporal context of the situation they address. Situational questions provide a hypothetical situation and ask applicants *what they would do if* they were in this situation. In contrast, past-behavior questions invite applicants to provide a past situation and ask *what they have done when* that situation happened. The former relies on applicants' intentions as a predictor of future behaviors, assuming that one does what one says (Latham et al., 1980). The latter presupposes that one's past behaviors will likely be the same in the future, therefore predicting future behaviors (Janz, 1989).

Behavioral questions have a good predictive validity, with past behavior questions having a higher one (Hartwell et al., 2019; Huffcutt et al., 2004; Taylor & Small, 2002). However, there is still some uncertainty regarding how well these questions truly capture the intended construct. While both question types measure the same construct (i.e., a specific competency) and their scores strongly correlate (Campion et al., 1997; Conway & Peneno, 1999), they show different relationships with other interview measures. Situational questions mainly correlate with job knowledge and cognitive ability, whereas past-behavior questions appear to correlate with job experience, job knowledge, social skills, and certain personality traits, such as extraversion (Conway & Peneno, 1999; Day & Carroll, 2003; Krajewski et al., 2006; Salgado & Moscoso, 2002).

To ensure the predictive validity of behavioral questions, recruiters should evaluate applicants' responses using pre-established behaviorally anchored rating scales (BARS)

that contain definitions and examples of typical responses (Roulin et al., 2012). BARS are indeed known to mitigate various biases in evaluations, among others, biases toward gender, sex, and race (Smith & Kendall, 1963).

In summary, behavioral interviews provide a structured framework for predicting applicants' future performance. However, this approach falls short of helping us to grasp interpersonal interactions, such as how applicants and recruiters influence each other through their behaviors. To fully understand applicants' responses — and how their behaviors can vary and influence interview outcomes — it is essential to consider the social dimension of the interview context.

1.2 Social Aspects of Job Interviews

Job interviews constitute social interactions where recruiters and applicants meet with a shared goal: exchange information to evaluate the potential for a lasting professional relationship (Bauer et al., 1998; Levashina et al., 2014). As in any social interaction, individuals seek to positively influence how they are perceived by presenting themselves in a favorable light to others (Goffman, 1956). To achieve this, both recruiter and applicants can adopt verbal and non-verbal behaviors to shape the desired self-image. For example, they can take care of their appearance and communication.

Interactions between recruiters and applicants are thus part of a dynamic process in which one's actions influence those of the other. This mutual influence not only shapes the progress of the interaction or the others' impressions but also the interview outcome. As in a chess tournament, each party adjusts its strategy in response to the other's moves, determining the flow of the interaction and the final decisions. In this light, I first explain what self-presentation is in the context of social interactions and then in job interviews (1.2.1), and review the different impression management strategies serving self-presentation that operate in job interviews (1.2.2).

1.2.1 Opportunities for Self-Presentation

Self-presentation in interaction is defined as "establishing, maintaining, or refining an image of the individual in the minds of others" (Baumeister, 1982, p. 3). Interestingly, Goffman (1956) uses a theater metaphor to explain this process in our everyday life. In his metaphor, individuals behave as actors playing a role, therefore adopting behaviors to create a desired image of oneself. A key idea here is that individuals have a *face*, that

is, a positive social value they aim to maintain or improve (Goffman, 1967). However, maintaining face hinges not just on the actor but also on the audience. Individuals seek to maintain their social image while helping others maintain theirs during interactions (Goffman, 1956, 1967).

In other words, self-presentation mobilize behaviors, consciously or unconsciously, that communicate information about the self to shape the desired image, real or false. These behaviors can be verbal — such as choice of words or intonation — and non-verbal behaviors — such as body language and physical appearance (Goffman, 1956). With this, the aims are to please others and to create a favorable impression on others (Baumeister, 1982). Depending on the context and norms, individual adopts specific strategies to achieve their goals.

Job interviews are social settings prone to self-presentation for both recruiters and applicants (Barrick et al., 2009). Recruiters aim to make the job and organization appear as positively as possible to attract applicants (Wilhelmy et al., 2016), while applicants can use various strategies with the aim to increase their hirability (Barrick et al., 2009). Indeed, recruiters' subjective impression of applicants affects interview outcomes (Stevens & Kristof, 1995). For example, choosing a professional outfit may help them manage how they are perceived. They can also adjust their speech fluency and word choice based on the cues received during the interaction while maintaining a confident posture, smile, and eye contact.

1.2.2 Impression Management in Job Interviews

A broader form of self-presentation tactics is impression management (IM), which is defined as a social influence process aimed at maximizing desired outcomes while minimizing undesirable ones (Leary & Kowalski, 1990; Schlenker, 1980). IM, therefore, involves a conscious or unconscious attempt to control the self-image projected in interaction by adopting verbal and non-verbal behaviors to influence the attitudes and behaviors of others (Bolino et al., 2008; Schlenker, 1980). IM occurs particularly in job interviews due to their high stakes nature (Rosenfeld, 1997). When engaging in IM, applicants shape their image and aim to convince recruiters that they are the best fit for the job (Barrick et al., 2009; Peck & Levashina, 2017; Proost et al., 2010; Van Iddekinge et al., 2007). Research has shown that applicants frequently use IM during interviews (e.g., Ellis et al., 2002, Levashina et al., 2014, Posthuma et al., 2002, Stevens & Kristof, 1995). In their study, Ellis et al. (2002) reported that most applicants, 97.5%, employed at least one form of IM during their interview.

IM strategies can be divided in different categories. One common distinction divides these strategies into assertive and defensive categories (Bolino et al., 2008). Assertive IM aims to proactively build and promote a positive self-image through behaviors like self-promotion to highlight one's accomplishments or ingratiation to evoke interpersonal liking; defensive IM seeks to repair one's image when it is threatened, often through excuses or justifications (Bolino et al., 2008; Stevens & Kristof, 1995). Assertive and defensive tactics can also be distinguished by the target: self or other. Self-focused tactics aim at highlighting one's strengths and achievements (e.g., self-promotion); other-focused tactics (i.e., aimed at recruiters or the organization) emphasize similarities or inspire sympathy in the target (e.g., ingratiation; Kacmar et al., 1992). While assertive IM is linked to positive outcomes, such as perceived job suitability and higher performance evaluation ratings, defensive IM is less effective, and its excessive use risks creating an impression of insecurity and lack of confidence (Bolino et al., 2008; Ellis et al., 2002; Stevens & Kristof, 1995). These assertive and defensive IM strategies can also be categorized as either honest strategies or their opposite, deceptive strategies: IM is considered honest if it reflects reality, but it is deemed deceptive when it distorts reality to make oneself appear better than one is (Bourdage et al., 2018; Levashina & Campion, 2007). Of note, applicants may engage in both honest and deceptive strategies within the same interview.

IM can also occur through controlling non-verbal behaviors (Schneider, 1981, Stevens & Kristoff, 1995). Applicants can adjust their gestures and so-called "immediacy" behaviors, including frequent eye contact with the recruiter, smiling, adjusting their body posture, or reducing interpersonal distance. Such IM can convey an impression of being confident, motivated, and competent and thus positively affects recruiters' evaluations (Burnett & Motowidlo, 1998; Gifford et al., 1985; Imada & Hakel, 1977; Tessler & Sushelsky, 1978). In contrast, applicants who cannot control their behaviors may create unfavorable impressions. For example, a person unable to manage stress and anxiety-related behaviors during the interview may be perceived as lacking confidence and warmth (Feiler & Powell, 2016; Finnerty et al., 2016; Vijay et al., 2021). However, non-verbal behaviors must form a consistent image with verbal behaviors, or there is a risk of being perceived as an imposter (Goffman, 1956).

Finally, because each applicant has a different background, there might also be differences in how they use IM tactics. These differences may come from cultural norms that dictate what behavior is appropriate or not.

1.3 Cultural Differences in Job Interviews

Culture encompasses a set of "values, attitudes, beliefs, and behavioral meanings shared by members of a society" (Manroop et al., 2013, p. 3516), shaping how individuals navigate situations like job interviews. As such, culture shapes not only what we do but also how we understand and respond to the behaviors of others around us. It is thus essential to consider cultural differences to understand how people behave in job interviews and how this affects others.

According to Hofstede (1984, 2010), culture can be understood through six dimensions: Members of a culture can be characterized by how they deal with social inequality (power distance), how they relate to the community (individualism), how they avoid uncertainty (uncertainty avoidance), whether the shared goals are masculine with work as a central aspect (masculinity), how they are committed to traditions (long-term orientation) and how they tolerate natural human drives (indulgence; Triandis, 1982).

These cultural dimensions set underlying social norms and rules that shape its members' verbal and non-verbal behaviors (Liu et al., 2022). As cultures differ on these dimensions, so do perceptions and expectations of 'normal' behaviors. For example, American culture views speaking quickly with short pauses as positive (Jaworski, 1993). In contrast, other cultures, such as Chinese and Japanese, favor speeches that include pauses (Endrass et al., 2008; Gao & Ting-Toomey, 1998). Maintaining eye contact is positively evaluated by Western Europeans but it is not the case for East Asian cultures (Uono & Hietanen, 2015). Moreover, some cultures use prominent and expansive spatial gestures, while others favor more compact and restricted gestures (Kita, 2009). Thus, one adopting behaviors that do not match others' cultural expectations is likely to be perceived negatively or, at least, identified as a foreigner.

Culture also influences how individuals behave and are evaluated in job interviews. For example, American culture tends to favor applicants who show enthusiasm in interviews more, whereas Chinese culture leans toward calmer applicants (Bencharit et al., 2019). Additionally, applicants' IM strategies for crafting a favorable self-image differ across cultures: Applicants from performance-oriented cultures are more likely to promote themselves and their achievements than those from cultures that place less value on performance (Arseneault & Roulin, 2021; Sandal et al., 2014). Hence, applicants' success in self-presentation in job interviews strongly hinges on the cultural environment, an important factor to consider in understanding job interview processes and outcomes.

Because job interviews entail both psychometric and social perspectives where self-presentation plays a key role but varies across cultures, this thesis focuses on past-behavior questions to investigate interview processes and outcomes. In the next chapter, I start examining how applicants respond to past behavior questions, specifically focusing on response framing and how this affects recruiters' perceptions and evaluations. Then, I continue deepening our understanding considering the role of culture and technology in job interviews, addressing applicants' response content, recruiters' perceptions and evaluations of performance and interview criterion validity (later in Chapter 3).

2 Responding to Past-Behavior Questions

> [P]ast-behavior questions invite a narrative response about the situation, actions undertaken by the candidate and related events that transpired. In other words, candidates are expected to respond to such questions by telling a story
>
> **Brosy et al. (2016, p. 373)**

When recruiters ask past-behavior questions, they invite applicants to recount relevant past work-related situations that showcase their mastery of competencies. Applicants thus have to recall a relevant situation quickly and respond to the question, describing it (e.g., the people involved and the problem) and explaining how they managed it.

Applicants should thus engage in *storytelling*, that is, narrate "a set of events related to a unique past episode characterized by a unity of time or action, which constituents are often linked with temporal markers" (Bangerter et al., 2014, p. 598). Ideally, they should follow the STAR mnemonic: start with a brief description of the problematic situation (S), detail their task and actions undertaken to solve the problems (TA), and mention the results obtained (R; Kessler, 2006). Moreover, applicants should think about how they want to frame their stories emotionally, depending on the impression they want to create. Do they emphasize the negative aspects of the situation, the positive ones, or mention both? The same story, but with different emotional framing will likely create different impressions on recruiters.

Nevertheless, applicants' storytelling is not about delivering a story to passive recruiters. Storytelling is a collaborative activity requiring the active engagement of both the narrator and the audience, respectively the applicant and the recruiters. Together, they co-create the narrative (Mandelbaum, 2013). To ensure a shared understanding of the story as it unfolds, the audience collaborates with the teller, providing either generic responses (e.g., "mhm", "yes") or specific responses (e.g., "No way he did that!") to indicate understanding and attention (Bavelas et al., 2000). When the audience is distracted and cannot engage in the co-creation, it diminishes the overall quality of the story (Bavelas et al., 2000). Therefore, applicants' storytelling heavily depends on recruiters' active listening.

Given that recruiters should cooperate in the storytelling activity and at the same time evaluate the response, storytelling responses to past-behavior questions merge the

social and psychometric aspects of job interviews. In this context, understanding applicants' responses requires examining both how their answers are structured and how they influence recruiters' evaluations. Therefore, I start by examining storytelling responses to past-behavior questions, considering their effectiveness and highlighting challenges for applicants (2.1). Building on this, I investigate the role of response framing on recruiters' evaluations (2.1.1), thereby presenting Study 1, which examines how emotional framing in applicants' narrative responses affects interview outcomes (2.1.2).

2.1 Storytelling Responses to Past-Behavior Questions

Storytelling responses are particularly effective for answering past behavior questions because they offer applicants a way to manage impressions for several reasons (Ralston et al., 2003). First, storytelling may appear as an honest account because it is difficult to fake (Bangerter et al., 2014). Second, it also allows applicants to present themselves positively and control how they are perceived. As the main character in their stories, applicants can attribute desirable characteristics and values to themselves (Silvester, 1997). For example, they can portray themselves as fully motivated to overcome challenges. Third, storytelling is more persuasive than simply listing facts (Krause & Rucker, 2020), as it captivates the audience in the vivid imagery of the characters (Winkler et al., 2022). When engaged, the audience is more receptive to the story's message, a phenomenon known as narrative transportation. This transportation can influence their attitudes by reducing their critical thinking (Van Laer et al., 2014).

However, telling stories on the fly during job interviews is challenging and cognitively demanding. Storytelling requires recalling a suitable past event, organizing the response, and delivering it adequately within a short time frame after recruiters' question (Brosy et al., 2020). This process proves difficult for many applicants, as they rarely deliver fully developed stories without assistance (Bangerter et al., 2014; Brosy et al., 2016, 2020). The few engaging in storytelling heavily rely on describing the situation, omitting to mention their actions and results that would serve to illustrate their competence (Bangerter et al., 2014). Instead, many resort to suboptimal narratives entailing broad descriptions of generic situations (i.e., pseudo-stories, "In this kind of situation, I usually stay nice and calm"). Under time pressure, they may also provide decontextualized assertions about their values and opinions ("I think that one should always be nice and calm") or self-description ("I am a calm person"). However, only narrative responses (i.e., stories and pseudo-stories) lead to higher hirability ratings (Bangerter et al., 2014).

Importantly, applicants' responses can be improved through different strategies. The

first strategy involves training before the interview. One approach is that a coach provides detailed feedback on the response, helping the applicant construct a corrected version of the narrative, which is then discussed within a training group (Ralston et al., 2003). A second approach consists of extended training with image-based intervention wherein applicants create images associated with their examples (Lin-Stephens et al., 2022). However, these procedures are time-consuming and difficult for applicants to implement (e.g., the lack of access to a coach). Alternatively, another strategy relies on the active participation of recruiters: They may use probe questions during the interview to help applicants improve the content of their answers (Brosy et al., 2020). While these strategies focus on enabling applicants to tell stories, we still understand little about what constitutes a compelling response and how it impacts recruiters. One starting point for the answer may be how responses are framed.

2.1.1 Emotional Framing in Narrative Responses

The framing (*how something is said*) of an experience can influence evaluation beyond *what is said*. Consider the following question: "Can you describe a challenging situation where you had to manage an angry client?" An applicant might modestly share an experience where an angry client approached the customer service desk and found an appropriate solution based on their discussion. Now, let us imagine a different telling where the applicant highlights the client's loud approach, shouting in front of others, and the complex negotiations leading to a solution, leaving the client smiling and grateful to the team. By emphasizing the negative aspects of the initial situation and the negotiation's complexity, ending with the client's great satisfaction, the applicant could provoke a different reaction from recruiters. The emotional elements in the narrative may greatly influence the audience's reaction.

The effect of emotional elements may stem from how they are introduced and fluctuate within the narrative. Indeed, narratives are characterized by their structure and their emotional patterns. On the one hand, narratives all share a common structure, including a beginning with the description of the context and protagonists (staging), a middle where actions unfold (plot progression) and a cognitive tension that builds in parallel to a peak just before the end, and finally, the narrative concludes (Boyd et al., 2020).

On the other hand, narratives also tend to revolve around core emotional trajectories. One study identified six emotional trajectories around which various narrations cluster (Reagan et al., 2016). The first trajectory is the constant rise of positive emotions, illustrated by the "rag-to-riches" story of a poor man who becomes wealthy. The second

shows a decline in positive emotions, illustrated by "Lady Susan", about an ambitious woman who marries below her standards. The third trajectory starts with a fall followed by a rise in positive emotions, featuring a protagonist who overcomes a significant setback (e.g., Man in a Hole). The fourth opens with a rise followed by a fall in positive emotions, where the protagonist starts on a high but plunges into despair (e.g., Icarus). The fifth is an emotional rollercoaster – a rise, fall, and rise again in positive emotions, exemplified by Cinderella. The sixth and final trajectory is the opposite of the rollercoaster with a fall, rise, and fall in positive emotions (e.g., Oedipus).

Interestingly, emotional trajectories contained in narratives encourage reactions and behavioral changes in the audience. For instance, studies in health communication have identified that when a message is emotionally framed with emotions shifting from a negative emotional beginning to a positive emotional ending (or vice versa), it can induce more behavioral changes in the audience than when the emotions do not fluctuate throughout the narrative (Alam & So, 2020; Carrera et al., 2008; Nabi, 2015). In the organizational context, similar indicators can be found when charismatic leaders manipulate emotions in their speeches to guide the audience toward desired change (Sy et al., 2018; Wasielewski, 1985).

Thus, the trajectory of positive and negative emotions in narratives may contribute to the persuasive power of stories. Narratives have a natural propensity to transport their audience by immersing them in evocative imagery conveyed by the protagonist (i.e., narrative transportation). However, the emotional trajectories may heighten transportation by fostering a deeper emotional connection with the narrative (Green & Appel, 2024; Green & Brock, 2000). As such, narrative responses incorporating emotional trajectories may increase recruiters' narrative transport, reducing their critical thinking and influencing their perceptions and thus evaluation of applicants' responses. Although it may affect interview outcomes and ultimately its validity, no research, to my knowledge, addressed that.

2.1.2 Study 1: Effect of Emotional Framing in Narrative Responses to Past-Behavior Questions on Interview Outcomes

In Study 1, we improved our understanding of what leads narrative responses to past-behavior questions to higher evaluations through the lens of emotional framing.

Previous research evidenced that emotional trajectories in narratives influence the audience's reactions and behaviors (Alam & So, 2020; Carrera et al., 2008). Some applicants

may, consciously or not, provide narrative responses with effective emotional framing to enhance their persuasive impact. Specifically, best-rated narrative responses would structure their emotional trajectories to show progress toward a goal by overcoming challenges, corresponding to an ascending positive emotional trajectory and a descending negative emotional trajectory. However, their role in shaping recruiters' evaluations during job interviews remains unexplored.

Considering the critical role of applicants' communication skills in interview validity (Huffcutt & Murphy, 2023), understanding the emotional trajectories in applicants' responses can provide insights into how their communication contributes to interview outcomes. To address this gap, we were thus interested in examining how positive and negative emotions fluctuate over the course of narrative responses to past-behavior questions (Research Question 1) and how positive and negative emotional trajectories within narrative responses affect recruiters' evaluations and emotional engagement with the responses (Research Question 2).

To answer our research questions, we used face-to-face (FTF) mock interviews that included three past-behavior questions that naturally elicited narrative responses. We used LIWC software (Boyd et al., 2022; Piolat et al., 2011) to extract positive and negative emotions from the transcripts of narrative responses (i.e., stories and pseudo-stories). This software categorizes words from a given text into positive and negative emotion categories and provides the percentage of words that fall into each category. Based on the video recordings, raters evaluated participants' competence using BARS anchored with examples and definitions. They also evaluated performance, engagement, and persuasiveness and one rater reported her emotional engagement with the narrative response. These latter evaluations were measured using 5-point scales and form more "general" and "subjective" measures compared to the BARS-based evaluations of competence.

Regarding the fluctuation of positive and negative emotions in narrative responses (Research Question 1), analyses revealed that, on average, the positive emotions significantly increased throughout the response forming a rising trajectory, while the quantity of negative emotions remained constant showing a flat trajectory. Although job interviews constitute formalized high-stakes social interaction, this finding shows that emotions are part of those interactions and may fluctuate to follow trajectories, as they would in more traditional narratives (Reagan et al., 2016).

Regarding the effect of emotional trajectories on raters' evaluations (Research Question 2), analyses showed that the emotions and their trajectories within responses influenced the raters' evaluations. A higher presence of negative emotions in the overall

narrative response, along with increasing negative emotions throughout the response, was linked to lower competence ratings; increasing negative emotions during the response also resulted in a decreased performance. This suggests that negative emotions may signal that the participant is not doing well. On the other hand, a higher presence of positive emotions in the overall narrative response resulted in a lower engagement. This may indicate that too much positivity signals a lack of involvement. However, increasing positive emotions throughout the response enhanced the persuasiveness ratings, signaling that the person is progressively solving problems.

Surprisingly, the rater's emotional engagement could not be explained by positive and negative emotions or their trajectories. Following the logic that narrative transportation is enhanced by emotional trajectories, making the message more persuasive (Alam & So, 2020; Carrera et al., 2008), one might expect that the rater's emotional engagement would be related to emotions in narrative responses. Our finding may be explained by the fact that, in this study, the rater evaluated their emotional engagement based on multimodal responses, specifically video-recorded narrative responses. Previous research linking the audience's emotional engagement, narrative transportation, and behavior changes used written narratives, which concentrated the emotional dimension solely on the words (Alam & So, 2020; Carrera et al., 2008). In contrast, our rater viewed video-recorded narratives that included both audio and visual cues. However, emotions can be conveyed beyond words through non-verbal and paraverbal cues, such as facial expressions and intonation variations (App et al., 2011; Simon-Thomas et al., 2009). Thus, the mere emotional valence of spoken words could not sufficiently explain the rater's emotional engagement.

Nevertheless, these results indicate that the emotional framing of a narrative response can influence recruiters' evaluations during an interview. Results are further discussed in the Discussion (Chapter 4).

3 Technology in Job Interviews

> Information technology has had widespread effects on almost every aspect of our society. [...] It has also had a profound impact on organizational processes, including those in Human Resource Management.
>
> **Stone et al. (2015, p. 216)**

Technological advances have transformed hiring practices by introducing new interview formats and tools that offer a range of unique possibilities. One format that has emerged recently and continues to gain popularity is the asynchronous video interview (AVI). In AVIs, recruiters upload interview questions to online platforms and applicants record their responses with their computer webcams at their convenience (Lukacik et al., 2022). However, AVIs deprive of real-time interaction, changing how applicants and recruiters connect and exchange information. This new form of social interaction raises questions about its impact on applicants' experiences, response behaviors such as storytelling, and on interview outcomes and validity.

In addition to AVIs, artificial intelligence (AI) has also been incorporated into the personnel selection process through tools designed to assist recruiters in their tasks. AI mimics human intelligence using machines (computers) to perform tasks that would typically require humans (Janiesch et al., 2021). Hence, AI models offer promise for developing tools that can assist in hiring decisions. At the same time, they could also help train applicants to respond appropriately to interview questions with automated analyses of their behaviors, thus enabling the provision of specific feedback.

Within the framework of this thesis, I investigate how technological advances in the form of AVIs and AIs impact job interview processes and outcomes. I first discuss AVIs, detailing what they are and how they differ from FTF interviews, including applicants' reactions and performance in such interviews (3.1). Building on Study 1, which demonstrated the effect of well-framed narrative responses on recruiters' evaluations, I examine how applicants' culture and the lack of real-time interaction in AVIs can affect applicants' storytelling responses and recruiters' evaluations. To this end, I compare applicants' responses to past-behavior questions, their interview performance, recruiters' evaluations of them and interview validity in FTF and AVIs (Studies 2 and 3; Section 3.2). Finally, I explore the potential of AI-powered platforms to assist in training applicants to respond to interview questions. Study 4 investigates the feasibility of analyzing response content

directly from interview audio recordings, which would bring us closer to AI models that could be implemented on platforms providing personalized feedback on responses (3.3).

3.1 Asynchronous Video Interviews (AVIs)

In recent years, organizations have manifested a growing interest in using AVIs (Ciphr, 2023; Griswold et al., 2022). HireVue, a leading company providing AVI services worldwide, reported that over 1 million interviews were completed on its platform during October 2021 only (Hirevue, 2021). This exponential interest may stem from their cost-effectiveness in rapidly interviewing applicants, adaptable scheduling releasing from the constraints to meet in-person or remotely, greater capacity to address a broader pool of applicants, and the ability to defer evaluations (Lukacik et al., 2022). Besides their practicability, AVIs also include standardization of some interview aspects, which helps increase reliability and validity (Lukacik et al., 2022): They present the same interview questions to all applicants and allow the same response time across participants. Moreover, AVIs prevent asking any other questions than those planned, which increases standardization and reduces biases (Levashina et al., 2014).

Besides deciding what question to ask, recruiters have to choose among various possible AVI designs, which are how the different features of the interview are configured. For instance, interview questions can be presented in different formats: text-based (questions written on-screen), avatar-based (an avatar asks the questions), or recorded videos (an interviewer asks the questions). As for the response, applicants can be (optionally) provided time to prepare their response and a specific time allotted to respond. Moreover, they may be allowed to review and re-record their responses (see Lukacik et al., 2022). Thus, AVIs can widely vary from one another due to recruiters' unique decisions for each feature and are likely to elicit different applicants' experiences and reactions.

Despite AVIs' practical benefits for both recruiters and applicants, they pose new challenges as they fundamentally alter the classical interview dynamics between applicants and recruiters. The one-way interaction deprives applicants of immediate feedback and non-verbal cues from recruiters. Applicants must then navigate the challenge of crafting an optimal self-presentation without being able to gauge the impression they are creating on recruiters. They must also monitor their responses without relying on the help or intervention of recruiters. Therefore, AVIs are likely to affect interview processes and outcomes compared to synchronous interview media (e.g., videoconference and face-to-face interviews).

To grasp the impact of this new interview format and its different designs, we first need to understand how this form of social interaction differs from traditional FTF interviews (3.1.1) and how applicants react and perform in such settings (3.1.2). These aspects are essential for further exploration of how the interview medium may shape interview processes, outcomes and validity (a focus in Chapter 3.2).

3.1.1 Media Richness and Social Presence in AVIs

The one-way interaction in AVIs deeply changes the dynamic compared to FTF interviews. Potosky's (2008) framework of four media attributes helps us understand how AVIs differ from traditional in-person interviews. The first of these attributes is interactivity, defined as the pace and fluidity of exchanges between the interlocutors. FTF interviews inherently support a high degree of interactivity between recruiters and applicants, whereas AVIs fall short in this regard. Second, a medium is characterized by its social bandwidth, which refers to the amount of verbal and non-verbal cues it can transmit. FTF interviews allow for the full range of verbal and non-verbal behaviors. However, AVIs limit applicants, as only the upper part of their body is visible on video (Van Iddekinge et al., 2006). Additionally, recruiters cannot transmit cues themselves (Daft & Lengel, 1986; Potosky, 2008). Third, the transparency of the medium refers to the extent to which interlocutors are unaware of the medium during the interaction. In AVIs, applicants are constantly reminded that the computer is an obstacle between them and the recruiters. Fourth and lastly, the medium can be characterized by a feeling of surveillance, which refers to the possibility that a third party can enter the conversation or monitor it. While in FTF interviews, it is impossible for someone to access the interview without being seen by all, AVIs can raise issues regarding the confidentiality of the video recording: Who will see the video or where their responses are stored may not always be clear.

AVI's media attributes may thus influence the feeling of social presence, defined as "the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationships" (Short et al., 1976, p. 65). In FTF interviews, recruiters are physically present, fostering a strong sense of social presence. In contrast, applicants in AVIs, who face only a computer screen, experience a reduced sense of social presence. This perceived presence, in turn, influences how engaged individuals become in the interaction. Moreover, lower perceived social presence may impact the perception of opportunities to perform, thus, the perceived fairness of interpersonal treatment in the interview (Gilliland, 1993).

3.1.2 Applicants' Reactions and Interview Outcomes in AVIs

In general, applicants' reactions to AVIs are more negative than to other computer-mediated interviews (e.g., videoconferencing) and FTF interviews. Indeed, they report lower social presence and fairness, less opportunity to perform, and higher stress in AVIs than in FTF interviews (Basch et al., 2020; Kleinlogel et al., 2023). When further investigating their reaction to different AVI designs, research evidenced that each recruiter's design decision creates a different experience for applicants. For instance, limited or no preparation time may be seen as unfair, while longer preparation enhances the perception of opportunity to perform and treatment fairness (Basch et al., 2021; Langer et al., 2017). Similarly, the format of interview questions can improve the perceived social presence. While there seem to be no differences between text-based and avatar-based questions in terms of perceived fairness and opportunity to perform, video-recorded questions enhance perceived social presence compared to text-based questions (Kleinlogel et al., 2023; Rizi & Roulin, 2024).

This is an important issue that organizations should consider while creating their AVIs. Applicants' negative reactions toward the selection process can impact organizational reputation and outcomes such as job acceptance and turnover intentions (Bauer et al., 1998; Chapman et al., 2005; Hausknecht et al., 2004; Smither et al., 1993). When applicants perceive a diminished social presence and fewer opportunities to showcase their abilities (Basch et al., 2020), this perception may hinder their motivation to fully engage in the interview process (Hausknecht et al., 2004). However, motivation to engage with a test is crucial for validity (Schmit & Ryan, 1992). A decrease in motivation during AVIs may lead recruiters to misinterpret the applicants' competencies.

As for applicants' performance in AVIs, it differs from other interview media. Applicants get lower performance ratings in videoconferences than in AVIs (Langer et al., 2017). However, one study comparing FTF interviews with AVIs found that applicants' performance ratings were comparable across both media when using the same evaluation procedure, eliminating preparation time in the asynchronous condition and providing unlimited response time (Kleinlogel et al., 2023). However, providing preparation time can help applicants perform better (Roulin, Wong, et al., 2023).

3.2 Effect of the Interview Medium and Culture on Behavioral Interview Processes and Outcomes

3.2.1 Study 2: Effects of Interview Medium and Culture on Applicant Storytelling, Disfluencies and Evaluations in Behavioral Interviews

In Study 2, we further investigated the effects of different AVI designs (text-based vs. avatar-based questions) on applicants' responses and recruiters' evaluations of them compared to FTF interviews.

Because AVIs deprive applicants of two-way interactions with recruiters, applicant storytelling responses will likely differ between FTF and AVIs. Indeed, storytelling is a collaborative process requiring an active audience (Bavelas et al., 2000). Therefore, we investigated the effect of the interview medium on applicants' storytelling responses to past-behavior questions (Research Question 1).

Another aspect worth considering in applicants' responses is their speech fluency because it relates to how applicants are perceived and evaluated by recruiters. Disfluencies in one's speech are common in spontaneous conversations and manifest as silent pauses, fillers ("uh", "um"), interruptions, and repetitions (Fox Tree, 1995). While previously considered disruptive, these disfluencies can aid listeners in understanding and getting insights into the speaker's cognitive processes, like one's hesitation (Brennan & Williams, 1995; Corley & Stewart, 2008; Smith & Clark, 1993). In job interviews, applicants are expected to respond quickly after recruiters' questions (Brosy et al., 2016). However, they may need time to prepare their answers or fix syntactic or lexical issues, thus having disfluent speech (Brennan, 2000). In turn, these disfluencies negatively affect recruiters' perceptions of applicants and their evaluations (Brosy et al., 2016). In AVIs, applicants lack recruiters' feedback and non-verbal cues available in FTF interviews. They are forced to monitor their responses without immediate external validation or correction while they strive to have a natural speech flow. Hence, applicants' speech fluency might differ between FTF and AVI interviews, which may ultimately affect interview outcomes. Because of this new challenge, we explored how the interview medium impacts applicants' disfluencies in behavioral interviews (Research Question 2), in the form of fillers and repetitions.

The constraints posed by AVIs may also affect how recruiters evaluate applicants' engagement and self-confidence. For instance, the limited social bandwidth and the lack of direct interaction with recruiters prevent gestures or immediacy behaviors like eye contact and smiling (Imada & Hakel, 1977). However, such behaviors have been shown to

enhance the perception of self-confidence and motivation during FTF interviews (Gifford et al., 1985; Tessler & Sushelsky, 1978). Yet, in the context of AVIs, the absence of these cues raises questions about how recruiters' evaluations compare to those in traditional FTF settings. Initial research suggests that applicants appear less stressed in avatar-based and text-based AVIs than in FTF interviews (Kleinlogel et al., 2023). Nonetheless, there remains a gap in our understanding of how recruiters evaluate engagement and self-confidence across different AVI formats versus FTF interviews. Therefore, we asked how the interview medium affects recruiters' evaluations of applicants' engagement and self-confidence (Research Question 3).

Since one's behavior is influenced by culture (Liu et al., 2022), it is also important to consider cultural differences when studying verbal behaviors such as storytelling and disfluencies in the context of job interviews. These behavioral differences across cultures could result in different recruiters' evaluations. What is more, technology-mediated interviews can influence behavior differently across cultures. For example, the limited social bandwidth of AVIs can hinder applicants from cultures that use expansive gestures, limiting their self-expression compared to FTF interviews. In addition, attitudes toward technology can affect interactions; some cultures will be more open to computer-mediated communication. For instance, cultures high on power distance are more open to the use of technology than those lower on that dimension (Jan et al., 2022). Furthermore, cultural dimensions such as uncertainty avoidance, indulgence and short-term orientation may shape applicants' reactions to AVIs (Griswold et al., 2022). Such differences in attitude and reaction toward technology may yield different behaviors. Therefore, we investigated the effect of culture on applicants' responses and recruiters' evaluations across interview media (Research Question 4).

To answer our four research questions, we used data from Kleinlogel et al. (2023) to build on their research. Their convenience sample consisted of male Swiss and Indian participants who underwent a mock interview, including two past-behavior questions and three other questions from various question types (e.g., presentation questions, their quality, why they should be hired). Participants were attributed to one of three conditions: FTF vs. avatar-based AVIs vs. text-based AVIs.

Analyses of storytelling responses to past-behavior questions (Research Questions 1 and 4) showed that participants generated more stories in AVIs (avatar- and text-based) than in FTF interviews. But when comparing storytelling responses across interview media, we found that the amount of STAR narrative elements was consistent. This means that participants provided storytelling responses more frequently in AVIs, but in terms

of content, their stories are similar to those in FTF interviews. Results also evidenced cultural differences, with Swiss participants providing more storytelling responses than their Indian counterparts. Also, Swiss participants provided more situational and result-related narrative elements in AVIs than Indian participants.

Regarding disfluencies in behavioral interviews (Research Questions 2 and 4), findings indicated that repetitions were more frequent in avatar-based AVIs than in FTF interviews and text-based AVIs. At the same time, fillers did not vary across the three interview formats. Moreover, there was no difference between Swiss and Indian participants for both types of disfluencies.

Analyses of evaluations of engagement and self-confidence (Research Questions 3 and 4) revealed that participants were evaluated as more engaged in FTF interviews than in AVIs (avatar- and text-based). Also, Indian participants were evaluated as more engaged and self-confident than their Swiss counterparts.

This study highlighted that interview medium and culture can both affect aspects of applicants' responses and recruiters' evaluations. Thus, hiring organizations should carefully consider the effects of using lower-rich media like AVIs and cultural differences in job interviews. I further discuss the results in the Discussion (Chapter 4).

3.2.2 Study 3: Responses to Past-Behavior Questions in Face-to-Face and Asynchronous Video Interviews: Storytelling, Interview Performance and Criterion Validity

Study 3 further explored applicants' responses and performance across different interview media (FTF interviews versus text-based AVIs). At the same time, this study extended knowledge on the criterion (concurrent) validity of past-behavior questions in AVIs.

Although Study 2 demonstrated that storytelling responses tend to be more frequent in AVIs, findings are limited due to the use of a male-only sample. Furthermore, the narrow focus on storytelling in Study 2 excluded other response types, such as pseudo-stories and decontextualized assertions (i.e., assertions about one's values and opinions, self-description, and justifications), which may also vary according to the interview medium. To address these limitations and provide a more detailed understanding of applicants' responses, we formulated Research Question 1: What effect does the interview medium have on applicants' responses to past-behavior questions?

Applicants' interview performance across media has also been a subject of investiga-

tion. Initial studies indicate that applicants perform worse in videoconferences than in FTF interviews (Melchers et al., 2021) and AVIs (Langer et al., 2017). However, only one study compared applicants' performance between FTF interviews and AVIs and found comparable performance when the same evaluation procedure was used and preparation time was eliminated (Kleinlogel et al., 2023). To provide additional evidence to ensure that performance is comparable between the two interview settings, we developed Research Question 2: What effect does the interview medium have on applicants' interview performance?

Applicants' negative reactions to AVIs may jeopardize behavioral interview validity. Perceiving less social presence reduces their sense of opportunity to perform, reducing their motivation (Basch et al., 2020; Hausknecht et al., 2004). However, validity relies on one's motivation to engage with a test (Schmit & Ryan, 1992). Low motivation in AVIs can lead to recruiters misinterpreting applicants' competencies. Furthermore, applicants' higher stress in AVIs could also affect the validity (Kleinlogel et al., 2023; Schneider et al., 2019). To address these concerns, one way to assess past-behavior questions' criterion validity is to test applicants' performance in a standardized work sample (Heimann et al., 2020; Krajewski et al., 2006). This allows to compare the performance under controlled conditions and relates it to the interview performance. We formulated Research Question 3: What effect does the interview medium have on the criterion validity of past-behavior questions as measured in an experimental work sample?

We conducted an experiment over two sessions. In the first session, participants completed a standardized work sample consisting of two role-plays. From this, we derived a measure of exercise performance. A week later, in the second session, they completed one mock interview, either in the FTF interview or AVI conditions. During the interviews, participants answered three past-behavior questions, including one about the previous week's work sample. This experimental design links the performance at the past-behavior question about the work sample with the exercise performance at the work sample, allowing for criterion validity computation. It also enables us to relate the overall interview performance with exercise performance for overall interview criterion validity.

Regarding the effect of the interview medium on applicants' responses (Research Question 1), analyses revealed that participants talked more in AVIs than in FTF interviews by a factor of about 1.67. However, they did not differ in the rate of narrative answers (i.e., stories and pseudo-stories) or decontextualized assertions about self-descriptions and justifications. The only difference was that participants shared about their values and opinions in AVIs more than in FTF interviews.

Regarding the effect of the interview medium on interview performance (Research Question 2), we found no significant difference between FTF interviews and AVIs, aligning with findings by Kleinlogel et al. (2023).

As for the effect of the interview medium on interview criterion validity (Research Question 3), our results showed that the overall interview performance (i.e., responses to the three past-behavior questions) predicted exercise performance as measured by the performance at the role-plays. The interview medium had no effect, suggesting that its validity also holds in AVIs. Moreover, the performance of the past behavior question specifically targeting the role-plays also predicted the performance of the role-plays, aligning with Liff et al.'s (2024) evidence for criterion (predictive) validity of past-behavior question in AVIs.

Results are discussed in the Discussion (Chapter 4).

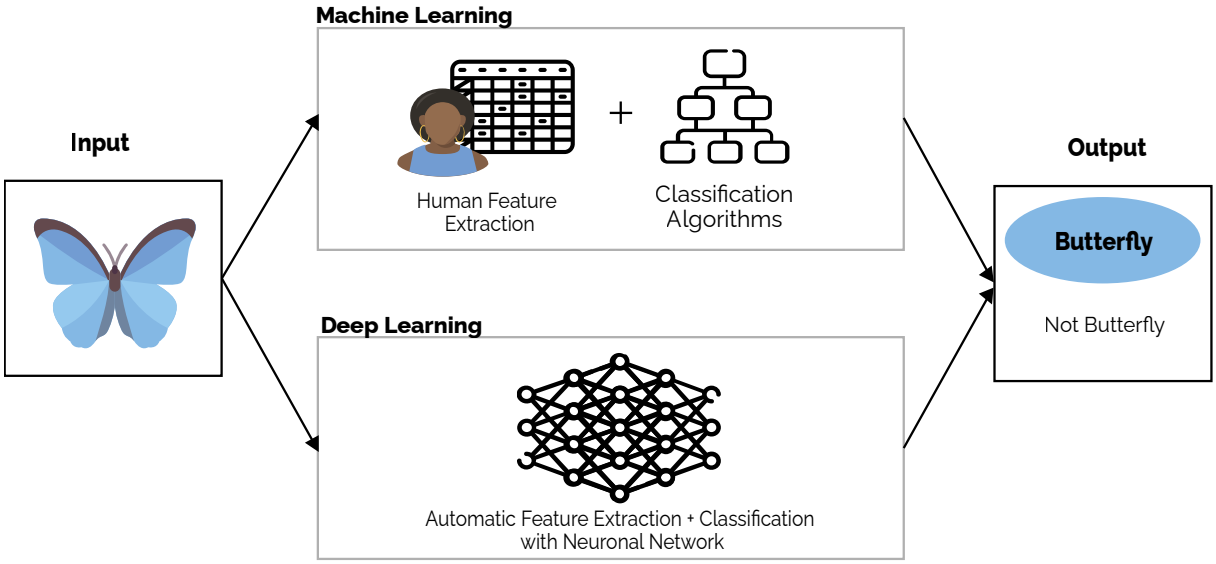
3.3 Artificial Intelligence (AI) Applications in Job Interviews

The application of artificial intelligence (AI) in job interviews is facilitated by AVIs. Its use is promising, as it can provide relevant information from recordings to support recruiters and applicants in this process. AI is a field of computer science that intends to create machines capable of performing tasks usually associated with human intelligence (Janiesch et al., 2021). A comprehensive review of AI's development and algorithms is beyond the scope of this thesis. Instead, I broadly introduce two subtypes of AI models, namely machine learning and deep learning models, to finally focus on their use in the context of job interviews.

Machine learning (ML) and deep learning (DL) are types of AI models that help computers understand the data and perform tasks by spotting patterns and relationships. ML uses algorithms to learn tasks by identifying relationships between human-provided annotations called *features* and outcomes called *labels* (Liem et al., 2018). Tasks include, among others, classifying data into categories (classification task) or predicting characteristics such as one's personality (regression task). DL models are more advanced types of ML models. They feature more complex structures that resemble (simplified) human brains with multiple layers of neurons (Khan et al., 2021). This structure enables them to automatically identify and extract relevant features from raw data, making human intervention unnecessary (see Figure 3 illustrating the difference between ML and DL).

To generate robust and accurate predictions, ML and DL models typically follow a two-step process: training and then evaluation (Liem et al., 2018). First, they are trained

Figure 3: Difference Between ML and DL on a Supervised Classification Task



to adapt to the data. This first step can typically be accomplished through supervised learning, where the model is provided with human-coded labels that serve as the ground truth that the model uses to learn, or unsupervised learning, where the model learns the rules by themselves (Janiesch et al., 2021; Liem et al., 2018). In the second phase, the trained model is tested and evaluated on previously unseen data to see how well it works with any newly presented data it will encounter when applied to practical use. Although key advantages of such models are their speed in processing large data sets and their ability to reveal relationships that humans may overlook (Khan et al., 2021), they highly depend on the quality of the training data; imprecise or biased data can lead to wrong decisions (Geiger et al., 2020; Liem et al., 2018).

In the context of job interviews, ML and DL models have gained interest for their potential to assist recruiters in decision-making. In this sense, various ML models have been developed to predict applicants' characteristics like personality (Hickman et al., 2022; Koutsoumpis et al., 2024; Rupasinghe et al., 2016) and outcomes related to performance or hirability ratings (Chen et al., 2017; Koutsoumpis et al., 2024; Naim et al., 2015; Nguyen et al., 2014). Based on video recordings, different behavioral cues – namely verbal (e.g., speech content), paraverbal (e.g., prosody, fluency), and non-verbal (e.g.,

facial expressions) – were extracted and input into the models. These models were then able to predict the desired characteristics satisfactorily. Moreover, the recent advances in DL have enhanced models that can now perform more powerful in-depth analyses of applicants' behaviors directly from AVI recordings to predict personality and hirability (Hemamou et al., 2019; Rahman et al., 2021; Suen et al., 2019).

Besides the opportunity to assist recruiters in hiring decisions, ML and DL models may be used to help applicants train for their future interviews. In the next subchapters, I explain why such automated training would be promising and review some developed training systems that provide feedback on behaviors (3.3.1). Then, I present Study 4, which explored DL models to analyze applicants' response type (storytelling and other response types) from interview audio recordings directly (3.3.2).

3.3.1 Machine Learning and Deep Learning for Applicants' Interview Training

To help applicants prepare for their future interviews, training platforms powered by AI could offer an alternative to conventional methods. Indeed, coaching with an instructor providing personalized feedback can help improve performance (Maurer & Solamon, 2006), but it may be difficult for some to access them. On the other hand, methods such as training with videos explaining the right behaviors can also help applicants (Roulin, Pham, & Bourdage, 2023), but do not provide personalized feedback, thus limiting individuals' improvement. In this light, training platforms equipped with automated analysis of behavioral cues would be able to provide both accessible training and personalized feedback to applicants, enabling them to improve their interview skills.

Such training systems have started to emerge. For instance, one application was developed for applicants to engage with a conversation avatar coach and get feedback on their non-verbal (e.g., smiling) and paraverbal (e.g., prosody) behaviors (Hoque et al., 2013). Another study incorporated automatic feedback regarding applicants' behaviors into serious games for job interview training (Gebhard et al., 2019). While interacting with the avatar coach, applicants get real-time feedback on seven non-verbal and paraverbal behaviors such as smiling, eye contact, body posture, arms crossed, nodding, voice volume, and voice energy. Additionally, Langer et al. (2016) found that providing real-time feedback on non-verbal behaviors during avatar-led job interviews improves interview performance and reduces interview anxiety.

However, these above training systems primarily focused on giving feedback on non-

verbal or paraverbal elements of applicants' responses, although verbal behavior, such as the response content, is a determinant factor explaining interview performance, often outweighing the importance of non-verbal or paraverbal behavior (Rasmussen, 1984; Riggio & Throckmorton, 1988). Therefore, developing models that provide feedback on the response content is important. To address this, Bangerter et al. (2023) attempted to identify storytelling and STAR narrative elements within applicants' transcribed responses in FTF interviews using various ML algorithms. Their results showed that it is feasible to predict applicants' storytelling responses. However, a limitation of their study is its dependency on human-generated transcripts. This makes it difficult to apply findings to platforms where only audio speech would be available. Moreover, it is impractical to manually transcribe audio recordings before conducting ML analyses in a scalable manner.

3.3.2 Study 4: Identifying Storytelling in Job Interviews Using Deep Learning

In Study 4, Mutian He and I (and, of course, all other co-authors) explored how more powerful DL models can help identify applicants' storytelling responses to past-behavior questions directly from interview audio recordings. Our goal was to pave the way for later audio-based models that could be implemented on training platforms to provide personalized feedback that improves applicants' storytelling skills.

Applicants seldom provide storytelling answers to behavioral questions, which enhance hirability, without prior training or help from recruiters' probe questions (Bangerter et al., 2014; Brosy et al., 2020; Roulin, Pham, & Bourdage, 2023). Interestingly, AI-driven systems in training platforms or software could offer tailored feedback to applicants based on their recorded answers. Building on Bangerter et al.'s (2023) work, we used the audio and transcript of mock behavioral interviews. Based on the transcripts, we divided responses to past-behavior questions into utterances, which are "a clause including a subject, verb, and object" (Bangerter et al., 2014, p. 598), and each utterance was classified for its utterance type, that is as belonging to a story vs. pseudo-story vs. decontextualized assertion vs. other. Utterances classified as stories and pseudo-stories were further classified as situation, task-action, or result-related elements, while decontextualized assertions were classified as values/opinions, self-description, and justification. This fine-grained labeling scheme thus contains 10 categories.

Before building DL models, we identified three issues to address. First, labels were unbalanced, with some very rare, making it difficult for the model to learn those labels.

We explored two alternative labeling schemes, one combining non-narrative labels (i.e., values/opinion, self-description, justification, and other), resulting in seven categories, and the other coarse-grained with only three categories, distinguishing story, pseudo-story, and other. Second, utterances were mostly short, which limited the context for models to learn and predict labels. We considered two techniques to enlarge the context: coalescence — merging adjacent utterances with the same labels, and expansion — providing a context of ∼20 seconds around the targeted utterance. Third, the model must handle audio recordings directly, as new data will not have human transcripts. Therefore, we used an automatic speech recognition (ASR) model to directly transcribe interviews from audio recordings.

We built a DL classification task that we developed in three steps, each step addressing the previously identified challenges. The first stage, termed *baseline assessment*, compared two types of models across the three labeling schemes: transcript-based models, which use manual transcriptions, and audio-based models, which use transcripts generated by ASR. This step helps us set a baseline for the audio-based models to measure their improvement in the later stages and compare them to the transcript-based models. The second step, termed *audio enrichment*, examined three ways to improve the performance of audio-based models by adding more information from audio recordings: 1) ASR mixing that combines human and ASR-generated transcripts during training, 2) audio inputs including paraverbal information extracted from the audio recordings, and 3) layer mixing which integrates information from different layers of audio data, such as higher layers for semantic meaning and lower layers for paralinguistic information. The third step, referred to as *context enlargement*, focused on expanding the context around an utterance using two techniques: coalescence and expansion techniques. These approaches were evaluated to determine their impact on model performance.

For the first step, the baseline assessment revealed two primary observations. Transcript-based models consistently outperformed audio-based models that did not incorporate audio enrichment or context enlargement. Also, simplifying the task by reducing the number of categories from ten to three enhanced the performance of both model types. This suggests that fewer labels result in more accurate predictions. The second step with audio enrichment showed that providing audio input helped in the coarser scenario with three broad categories. Conversely, ASR mixing performed better for scenarios with ten categories that would be more relevant in case of providing feedback on applicants' responses. However, layer mixing did not contribute much to performance improvements and brought considerable computational costs. Finally, for the third step with context enlargement, coalescence improved the performance of all audio-based models, though

context expansion proved to be the most effective, surpassing transcript-based models.

Results are further discussed in the Discussion (Chapter 4).

4 Discussion

> We live on an island surrounded by a sea of ignorance. As our island of knowledge grows, so does the shore of our ignorance.
>
> **John Archibald Wheeler**

By considering the behavioral job interview as a psychometric instrument and a social interaction, I aimed to extend our knowledge of what constitutes an effective response, how external factors such as the interview format and culture affect interview processes and outcomes, and how technological advancement may help train applicants efficiently. This thesis explored applicant narrative responses through four key dimensions: emotional framing in narratives (Study 1), cultural influences on responses and evaluations (Study 2), the impact of interview medium on responses, evaluations, and interview validity (Studies 2 and 3), and the use of advanced AI models to analyze response content (Study 4).

First, I explored the emotional framing of narrative responses to behavioral questions and their effects on evaluators (Study 1). I demonstrated that emotional trajectories within narrative responses generally included a rise in positive emotions during the response, while negative emotions stayed consistent. Moreover, both the total amount of emotions expressed and their progression throughout the response influenced evaluations. While negative emotions lowered competence and performance ratings, positive emotions had more nuanced effects: rising positive emotions positively affected evaluations of persuasiveness, but excessive amounts of positive emotions within the response lowered evaluations of engagement. This study highlights that not only does the response type (e.g., stories, pseudo-stories) lead to higher recruiters' ratings (Bangerter et al., 2014), but the response framing also contributes to these outcomes. Gaining insight into how recruiters evaluate applicants' responses thus requires considering applicants' communication more holistically.

Building on Kleinlogel et al.'s (2023) work, I then explored the effects of culture and interview medium (avatar-based and text-based AVIs versus FTF interviews) on applicant responses, focusing on storytelling, disfluencies, and recruiter evaluations (Study 2). Regarding the medium, I found that storytelling responses were more frequent in AVI settings, but the number of STAR narrative elements did not differ from FTF interviews. Also, there were more repetitions in avatar-based AVIs than in other interview settings, but fillers did not differ across interview media. Finally, I found that evaluations of

engagement were higher in FTF interviews, although evaluations of self-confidence did not vary by medium. Concerning the effect of culture, it affected the production of stories and recruiters' evaluations but not disfluencies: Swiss participants produced more stories, including more STAR elements, than Indian participants. Indian participants, however, were evaluated as more motivated and self-confident. These findings underscore the importance of considering both medium (and AVI design) and culture, as they play a role in interview processes and outcomes.

Third, I employed an experimental design to deepen our understanding of applicants' responses and performance in behavioral interviews across media (FTF interviews versus text-based AVIs; Study 3). Therefore, I focused on (1) comparing response types beyond storytelling, including pseudo-stories, decontextualized assertions such as values/opinions, self-descriptions, and justifications, (2) comparing performance across interview formats using BARS, and (3) evaluating the criterion validity of behavioral interviews. I evidenced that response content was largely consistent across media, except that participants shared more values and opinions in AVIs. In line with this, interview performance ratings did not differ by medium. Criterion validity remained across media, showing that interview performance in both interview formats predicted exercise performance effectively. These results suggest that AVIs with designs similar to FTF interviews can be a fair and valid interview method.

Finally, Mutian and I explored the feasibility of developing DL models to directly identify the content of behavioral responses from audio recordings (Study 4). We developed a classification DL pipeline in three stages. First, we compared models based on human-generated versus ASR-generated transcripts from audio recordings under three label scenarios. We found that audio-based models initially underperformed compared to transcript-based models. Second, we tried different techniques to enhance audio-based model performance: ASR mixing was particularly effective in scenarios with many categories to predict, while providing paraverbal audio information was beneficial in scenarios with few categories. Finally, we enlarged the context provided to the model and found that the audio-based model performance significantly improved, particularly through expansion, surpassing transcription-based models.

Since the limitations of each study have been discussed individually in the manuscripts, the following subchapters will first outline the main findings of this thesis and their implications for job interviews. I begin with the interview processes, specifically focusing on applicants' storytelling responses (4.1). Next, I continue discussing the findings on interview outcomes, that is, recruiters' evaluations along with the criterion validity of

behavioral interviews (4.2). I then conclude with the limitations and avenues for future research (4.3).

4.1 Factors Influencing Applicants' Storytelling Responses to Past-Behavior Questions and Opportunities for Improvement

Storytelling is optimal for applicants to respond to past behavior questions (Bangerter et al., 2014; Stevens & Kristof, 1995). On the one hand, they enable recruiters to assess their skills based on concrete experiences (Campion et al., 1997). On the other hand, they can lead to successful impression management: structured and credible accounts may create a perception of competence and increase the chances of being hired (Bangerter et al., 2014; Kessler, 2006; Stevens & Kristof, 1995). Although some criteria have been proposed to ensure effective storytelling (e.g., consistency, Ralston et al., 2003), little is known about how emotional framing plays a role in the effectiveness of narrative responses and in this thesis, I addressed this gap.

Study 1 revealed that the effectiveness of a story may depend not only on its structure and criteria like consistency but also on how it is framed. In particular, inappropriate use of negative emotions in the narrative may suggest to recruiters a lack of ability to handle problematic situations, leading to lower competence and performance evaluations. Conversely, positive emotions, as they progress gradually through the narrative, can reinforce the persuasive side of narratives by signaling a successful resolution to the situation. These results align with literature indicating that emotional trajectories in narratives contribute to audience reactions and behaviors (Alam & So, 2020; Carrera et al., 2008; Nabi, 2015; Sy et al., 2018; Wasielewski, 1985). This can be explained by the fact that emotional trajectories facilitate narrative transport, reinforcing the emotional link between the audience and the story and thus making them more receptive to the conveyed message (Green & Appel, 2024; Green & Brock, 2000). Assuming that narrative responses are honest and credible accounts of the applicant's skills (Bangerter et al., 2014), appropriate emotional framing can amplify its persuasive effects on recruiters and their chances of being hired. While more research is needed to identify and understand the constitutive factors in narrative performance, our findings provide initial evidence to Huffcutt & Murphy's (2023) call for investigating the role of applicants' communication skills in recruiters' evaluations and interview validity.

However, producing effective responses to past-behavior questions may depend not only on applicants' ability to create a well-framed narrative but also on how such be-

havior is perceived as appropriate within a cultural context. Indeed, the strategies used to manage impressions depend on what the culture values (Arseneault & Roulin, 2021; Sandal et al., 2014). This could thus imply that applicants' response content may also differ from one culture to another.

Study 2 revealed that applicants' cultural backgrounds indeed influenced their propensity to provide storytelling responses, as well as the content of these stories. The differences can be explained by the fact that the norms underlying social interaction differ from one culture to another (Liu et al., 2022). These norms may shape applicants' expectations of the questions that will be asked and their perception of what is appropriate and optimal to answer. For example, some cultures (e.g., Taiwan, the US) favor questions about personal values and opinions, while others do not (e.g., Belgium, Russia; Posthuma et al., 2014). Thus, applicants from a culture that is less familiar with past-behavior questions may be less aware that it is appropriate to answer them in narrative form. These results highlight that the production of narrative responses is not only a reflection of individual abilities to produce effective narratives but also of one's culture.

Beyond the impact of culture, studies 2 and 3 explored the impact of different interview formats on storytelling responses. In AVIs, applicants have to manage their responses without real-time feedback from recruiters (Lukacik et al., 2022), which can affect their behavior and lead to responses that differ from those in FTF. Our results concerning the production of stories are mixed: Study 2 suggests that storytelling responses are more frequent in AVIs than in FTFs, while Study 3 suggests no difference. However, both studies suggest that the content of applicants' responses is mostly comparable across different interview formats. I see two potential reasons for the different results regarding the frequency of stories across interview media. First, Study 1 included a diverse sample of Indian and Swiss participants, whereas Study 3 only involved Swiss undergraduates. Second, Study 2 used both avatar-based and text-question AVIs, whereas Study 3 solely used text-based AVIs. Thus, the combination of different cultures and AVI designs in Study 2 may have led to different storytelling response frequencies in AVIs.

Interestingly, I evidenced the subtle effects of different AVI designs on applicants' responses. Study 2 showed higher repetitions in avatar-based interviews. Given that disfluencies can indicate the cognitive load of the speaker (Smith & Clark, 1993; Corley & Stewart, 2008), this suggests that this AVI design brings an additional cognitive load: not only do they have to manage their response, but also deal with this unusual and strange form of interaction with an avatar (Brosy et al., 2016; Langer & König, 2018). In Study 3, responses in AVIs were 1.67 times longer and included more value- and opinion-related

statements, in contrast to Study 2, where response lengths were comparable across media. Again, the difference in response lengths across interview media between the two studies may stem from the AVI designs: AVIs in Study 2 had no response time limit and no on-screen timer, but AVIs in Study 3 had a visible timer on-screen indicating a five-minute response time limit. This on-screen timer may have encouraged participants to fill the remaining time with additional information (Orji et al., 2024). The content of these extended responses can be linked to narrative dynamics, where narrators want through their narratives to perform a social action, such as accounting for their behavior (Mandelbaum, 2013). In this way, narrators reveal the moral of the story at the end of their narration; if the audience does not respond, the narrators will repeat or emphasize this moral (Bavelas et al., 2000). Therefore, having a ongoing timer in an AVI could play the role of an unresponsive audience and encourage participants to clarify their stories by clearly articulating the moral, thereby mentioning their values and opinions at the end of their response. As such, AVI interview designs appear to provide a framework for applicants' response behaviors, so that when they differ, so do responses. For example, designs offering greater richness facilitate impression management behaviors (Rizi & Roulin, 2024).

However, producing rich and complete stories in the stressful context of a high-stakes interview remains a cognitive challenge for many applicants (Bangerter et al., 2014; Brosy et al., 2020). There exist various effective training approaches to help train applicants to storytelling responses (Lin-Stephens et al., 2022; Ralston et al., 2003; Roulin, Pham, & Bourdage, 2023). However, these approaches have significant limitations: they are often time-consuming and difficult to access for a wide range of applicants or they may not be personalized. With the idea of creating online training platforms, Bangerter et al. (2023) explored the feasibility of providing personalized feedback on the response content and have shown that it is possible to use ML to identify stories and their narrative elements. However, one limit of their approach is that it relies on human-generated transcripts, which limits its integration into automated training platforms.

Study 4 explored how DL techniques may help to overcome this limitation. We demonstrated that it is possible to automatically identify the response content directly from audio recordings, without the need for human transcription. Furthermore, our results have shown that DL algorithms can detect both storytelling and suboptimal responses, such as pseudo-stories and decontextualized assertions. However, to guarantee reliable results, it is essential to select techniques adapted to the specific task. In providing detailed feedback on narratives, it is important to use audio enrichment techniques capable of handling many labels. Furthermore, including a broader semantic context around

the target proved crucial, although it adds significant computational complexity. This complexity, which is often quadratic with the addition of contextual information, is a major constraint for current models (Vaswani et al., 2017). These results advance the possibility of automated coaching platforms that offer personalized feedback to applicants. However, implementing these systems will involve addressing challenges related to model complexity and computational costs. Therefore, Study 4 calls for more research to develop systems for interview response coaching that are accessible to a wide range of applicants.

Our studies suggest practical implications for both applicants and organizations. Firstly, organizations engaging in international selection need to take culture into account, both when developing interview questions and when assessing applicants, as our results suggest differences between Indian and Swiss individuals in the way they respond to questions. Secondly, applicants' responses do not differ significantly from one interview medium to another, although they do talk more about their values and opinions at the end of their AVI responses (Orji et al., 2024). However, this may depend on how the AVIs are configured (e.g., the time allotted for response, see 4.2). It is therefore important that organizations carefully consider the design of their AVIs and we reinforce previous advice to use the same interview medium (and design) for all applicants in a same selection procedure (Langer et al., 2017; Melchers et al., 2021).

4.2 Recruiters' Evaluations in Job Interviews: Challenges and Perspectives for the Validity of Behavioral Interviews

Behavioral interviewing is a best practice for predicting applicant job performance. Besides the psychometric advantages of this structured interview (Campion et al., 1997; Roulin et al., 2012), it entails a social dimension, offering applicants and recruiters the opportunity to shape each other's impressions (Barrick et al., 2009; Bolino et al., 2008). However, the social interaction unfolding during interviews depends highly on the context in which they occur. Two contextual factors warrant attention due to their influence on applicant behaviors. First, culture defines norms and expectations for social interactions (Liu et al., 2022), leading to different ways of managing impressions (Arseneault & Roulin, 2023; Sandal et al., 2014). Second, AVIs change job interview dynamics by eliminating real-time interactions. Thus, these two factors are likely to affect applicant behaviors and recruiters' evaluations.

In Study 2, I showed that recruiters' evaluations of applicants' self-confidence and engagement varied according to applicants' cultural background and the interview medium

used. These results suggest that cultural norms impact how applicants express engagement and demonstrate confidence. Of note, the raters of our Indian participants were Swiss, so we cannot exclude the possibility of cultural bias in our results (Arseneault & Roulin, 2023). Nevertheless, I have highlighted that applicants are evaluated as more engaged in FTF interviews than in AVIs. This difference can be explained by the greater co-presence and interactivity of FTF interviews (Basch et al., 2020). Higher interaction could not only foster rapport-building but also enhance their engagement. These results revealed that contextual factors, such as the interview medium and culture, influence applicants' behaviors and, consequently, recruiters' evaluations. However, if applicants' behaviors differ, it is also possible that evaluations of their performance vary, which could affect the validity of behavioral interviews.

In Study 3, I showed that interview performance did not differ between FTF and AVI interviews. This may be explained by the fact that, in Study 3, the content of applicants' responses was similar between media. Our results align with those of Kleinlogel et al. (2023), suggesting that AVIs can be a fair interview method. This may reassure applicants concerned about their lower chances of performing in AVIs (Basch et al., 2020; Kleinlogel et al., 2023). Moreover, I demonstrated that responses to past-behavior questions, and more broadly behavioral interviews, predict applicants' performance in an exercise, regardless of the medium. These results therefore indicate that the benefits of behavioral interviews in terms of criterion validity appear to extend to AVIs, confirming initial research (Liff et al., 2024). However, other AVI designs may affect applicant performance and play a role in interview validity. In studies 2 and 3, AVI designs were developed to be as comparable as possible with FTF interviews, which may not always be the case in practice. For instance, participants had no preparation time nor response time limit in Study 2, while in Study 3 they had 20 second preparation time and 5 minutes response time limit. This differs from Australian organizations that provide on average 30 second preparation time and 2 minutes response time limit (Dunlop et al., 2022). However, designs less similar to FTF interviews could lead to different performances. Current research shows that AVI design is indeed linked to applicant performance.

Certain AVI designs, such as preparation time or the possibility of re-recording responses, also seem to have an impact on applicants' responses and performance. Indeed, these possibilities enable applicants to perform better (Basch et al., 2021; Lukacik & Bourdage, 2024; Roulin, Wong, et al., 2023). But this raises a question: does this improvement reflect the applicants' competence or simply the fact that they have time to think about what to say and how to sound better? While Lukacik and Bourdage (2024) as well as Basch et al. (2021) found that longer preparation time favored honest IM without

increasing deceptive IM, Roulin, Wong, et al. (2023) observed a slight reduction in honest IM but no significant impact on deceptive IM. Regarding reregistration options, Lukacik and Bourdage (2024) found that these options improved both honest and deceptive IM, while Roulin, Wong, et al. (2023) reported a reduction in deceptive IM use but no significant change in honest IM. Furthermore, Orji & Bangerter (2024) tested three different response time limits (1 min versus 3 min versus 5 min) and found that performance was significantly better when the response time limit was three minutes.

Faced with all these design possibilities, it is now important to understand what AVI designs lead to performance that best reflects applicants' competence. With this in mind, future studies should continue to explore the effects of different designs on applicant performance and the criterion validity of interviews.

4.3 Limitations and Future Research

This thesis has some limitations and opens up several perspectives for future research. One significant limitation is that all our studies were conducted in laboratory settings. Although this experimental approach allowed us to rigorously control variables and examine causal relationships (Bless & Burger, 2016; Falk & Heckman, 2009), it does not fully reflect the social dynamics of a real interview. In particular, emotional and social stakes, such as the pressure associated with obtaining a position or power dynamics, were absent in our studies despite the incentive for participants to invest themselves through performance-based remuneration. Also, Study 4 relied on high-quality audio recordings, which often exceed real-life conditions where background noise or other technical interferences may occur.

A second limitation lies in the narrow operationalization of some variables from our studies. First, in Study 2, our exploration of cultural differences was restricted to a comparison between Swiss and Indian individuals. Although these groups differ on certain cultural dimensions (see The Culture Factor Group, 2024), other cultures may differ even more resulting in greater differences in how applicants respond and are evaluated. Consequently, our results must be interpreted with caution and cannot be generalized to other cultural contexts without further study. Second, Studies 2 and 3 mainly explored the differences between FTF interviews and AVIs, using a fixed parameter set-up that closely approximated that of FTF interviews. However, AVIs can vary considerably in terms of design (preparation time, possibility of re-recording, automated assessment; Lukacik et al., 2022). Although the examination of AVI designs was beyond the scope of this thesis, our results may not fully generalize to all AVIs. Finally, our studies focused on the story-

telling aspect of responses, helping us understand how responses can differ based on the emotional tone set (Study 1), culture (Study 2), or medium (Studies 2 and 3). However, we mainly relied on annotations from transcripted words. It is important to remember that applicants' responses during job interviews are inherently multimodal (except when using written response), meaning that words go hand in hand with other behavioral cues, such as non-verbal signals like gestures and facial expressions, as well as paraverbal elements such as intonation and pauses. These accompanying behaviors contribute to the construction of social interactions and can significantly affect how recruiters evaluate applicants. For instance, gestures can help organize important storytelling elements and increase clarity, making it easier for listeners to engage with the narrative (Jacobs & Garnham, 2007; Sekine & Kita, 2017).

Building on the limitations of this thesis and its studies, future research should first try to replicate the findings in real-life settings and across different job positions. Including a wider range of cultures and AVI designs would also help understand how applicants' responses may vary. Expanding the range of cultural backgrounds among study participants could give a deeper perspective on how cultural differences influence applicants' behaviors in job interviews. With increasing globalization, it is important to ensure that interview methods are fair for any applicant. Further, future studies could explore how different AVI parameters like preparation time, re-record options and response time limit influence applicants' storytelling responses. Enough preparation times may give applicants the possibility to find the most relevant situation and organize their thoughts on how they want to deliver their story. This may resolve the cognitive challenge associated with providing a response shortly after the question (Brosy et al., 2016) and help them showcase their competencies appropriately. Also, re-recording options might give them the opportunity to re-record if they felt that their response was lacking important narrative elements. Finding the optimal response time limit (not too long, not too short; Orji & Bangerter, 2024) would give enough time for applicants to unfold the story without risking extending their response with information such as values and opinions that could potentially harm their performance (Orji et al., 2024). Lastly, research would benefit from understanding more about applicants' communication and its impact on interview outcomes and validity, as proposed by Huffcutt and Murphy (2023). This would imply looking into how different behaviors—like verbal, non-verbal, and paraverbal—interact to form effective responses and how such combinations of behaviors affect evaluations.

To sum up, our studies have helped us gain an understanding of interview processes and their outcomes in various cultural and interview settings. Moving forward, future research should seek a more comprehensive view of how applicants respond in job inter-

views, considering a wider range of cultures, AVI designs, across job positions and with a more holistic approach to responses.

Conclusion

Job interviews are inherently social interactions where recruiters and applicants meet with a shared goal: exchange information to evaluate the potential for a lasting professional relationship (Bauer et al., 1998; Levashina et al., 2014). Previous studies mainly focused on psychometric aspects of job interviews (e.g., Campion et al., 1997; Huffcutt & Arthur, 1994; Hunter & Hunter, 1984; Schmidt & Hunter, 1998) and applicants' impression management (e.g., Bolino et al., 2008, Ellis et al., 2002, Stevens & Kristof, 1995). However, job interviews have recently changed a lot with technological advances. This thesis adds knowledge about how storytelling responses, recruiters' evaluations and interview validity compare across interview settings and cultures. It also shows how AI can assist applicants in this process. Still, more research is needed to understand interview processes and outcomes across AVI designs. Advancing this knowledge will not only enrich the field of personnel selection but also pave the way for inclusive and innovative hiring practices.

Declaration on the Use of AI Tools

In this thesis, I declare that I have used the following tools to correct my English: Grammarly (https://app.grammarly.com), DeepL (Deepl SE, https://www.deepl.com/translator), ChatGPT 4o (Open AI, https://chat.openai.com/chat). This includes checks regarding grammar, spelling, clarity of wording, translation of some of my ideas from French.

References

Alam, N., & So, J. (2020). Contributions of emotional flow in narrative persuasion: An empirical test of the emotional flow framework. *Communication Quarterly*, *68*(2), 161–182. https://doi.org/10.1080/01463373.2020.1725079

App, B., McIntosh, D. N., Reed, C. L., & Hertenstein, M. J. (2011). Nonverbal channel use in communication of emotion: How may depend on why. *Emotion*, *11*(3), 603–617. https://doi.org/10.1037/a0023164

Arseneault, R., & Roulin, N. (2021). A theoretical model of cross-cultural impression management in employment interviews. *International Journal of Selection and Assessment*, *29*(3-4), 352–366. https://doi.org/10.1111/ijsa.12348

Arseneault, R., & Roulin, N. (2023). Examining discrimination in asynchronous video interviews: Does cultural distance based on country-of-origin matter? *Applied Psychology*, *73*(1), 185–214. https://doi.org/10.1111/apps.12471

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research 1. *Personnel Psychology*, *35*(2), 281–322. https://doi.org/10.1111/j.1744-6570.1982.tb02197.x

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. *Journal of Business and Psychology*, *29*(4), 593–604. https://doi.org/10.1007/s10869-014-9350-0

Bangerter, A., Mayor, E., Muralidhar, S., Kleinlogel, E. P., Gatica-Perez, D., & Schmid Mast, M. (2023). Automatic identification of storytelling responses to past-behavior interview questions via lhine learning. *International Journal of Selection and Assessment*, *31*(3), 376–387. https://doi.org/10.1111/ijsa.12428

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (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*(6), 1394–1411. https://doi.org/10.1037/a0016532

Basch, J. M., Brenner, F., Melchers, K. G., Krumm, S., Dräger, L., Herzer, H., & Schuwerk, E. (2021). A good thing takes time: The role of preparation time in asynchronous video interviews. *International Journal of Selection and Assessment*, *29*(3-4), 378–392. https://doi.org/10.1111/ijsa.12341

Basch, J. M., Melchers, K. G., Kegelmann, J., & Lieb, L. (2020). Smile for the camera! The role of social presence and impression management in perceptions of technology-mediated interviews. *Journal of Managerial Psychology*, *35*(4), 285–299. https://doi.org/10.1108/JMP-09-2018-0398

Bauer, T. N., Maertz Jr, C. P., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology*, *83*(6), 892–903. https://doi.org/10.1037/0021-9010.83.6.892

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, *91*(1), 3–26. https://doi.org/10.1037/0033-2909.91.1.3

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*(6), 941–952. https://doi.org/10.1037/0022-3514.79.6.941

Bencharit, L. Z., Ho, Y. W., Fung, H. H., Yeung, D. Y., Stephens, N. M., Romero-Canyas, R., & Tsai, J. L. (2019). Should job applicants be excited or calm? The role of culture and ideal affect in employment settings. *Emotion*, *19*(3), 377–401. https://doi.org/10.1037/emo0000444

Bless, H., & Burger, A. M. (2016). A closer look at social psychologists' silver bullet: Inevitable and evitable side effects of the experimental approach. *Perspectives on Psychological Science*, *11*(2), 296–308. https://doi.org/10.1177/1745691615621278

Bolino, M. C., Kacmar, K. M., Turnley, W. H., & Gilstrap, J. B. (2008). A multi-level review of impression management motives and behaviors. *Journal of Management*, *34*(6), 1080–1109. https://doi.org/10.1177/0149206308324322

Bourdage, J. S., Roulin, N., & Tarraf, R. (2018). "I (might be) just that good": Honest and deceptive impression management in employment interviews. *Personnel Psychology*, *71*(4), 597–632. https://doi.org/10.1111/peps.12285

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22* (Computer software). Austin, TX: University of Texas at Austin. https://www.liwc.app

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, *6*(32), eaba2196. https://doi.org/10.1126/sciadv.aba2196

Brennan, S. E. (2000). Invited talk: Processes that shape conversation and their implications for computational linguistics. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 1–8.

Brennan, S. E., & Williams, M. (1995). The feeling of another' s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383–398. https://doi.org/10.1006/jmla.1995.1017

Brosy, J., Bangerter, A., & Mayor, E. (2016). Disfluent responses to job interview questions and what they entail. *Discourse Processes*, *53*(5-6), 371–391. https://doi.org/10.1080/0163853X.2016.1150769

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. *European Journal of Work and Organizational Psychology*, *29*(3), 330–343. https://doi.org/10.1080/1359432X.2019.1704265

Burnett, J. R., & Motowidlo, S. J. (1998). Relations between different sources of information in the structured selection interview. *Personnel Psychology*, *51*(4), 963–983. https://doi.org/10.1111/j.1744-6570.1998.tb00747.x

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655–702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Carrera, P., Caballero, A., & Muñoz, D. (2008). Comparing the effects of negative and mixed emotional messages on predicted occasional excessive drinking. *Substance Abuse: Research and Treatment*, *1*, 1–7. https://doi.org/10.1177/117822180800100001

Cascio, W. F., & Aguinis, H. (2014). *Applied psychology in human resource management* (6th ed.). New Jersey: Pearson Education.

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology*, *90*(5), 928–944. https://doi.org/10.1037/0021-9010.90.5.928

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, M. E. (2017). Automated video interview judgment on a large-sized corpus collected online. *2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 504–509. https://doi.org/10.1109/ACII.2017.8273646

Ciphr. (2023, July 6). *Are video interviews the future of recruitment?* Linkedin. https://www.linkedin.com/pulse/video-interviews-future-recruitment-ciphr/

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*(5), 565–579. https://doi.org/10.1037/0021-9010.80.5.565

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology*, *13*(4), 485–506. https://doi.org/10.1023/A:1022914803347

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589–602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, *32*(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554

Day, A. L., & Carroll, S. A. (2003). Situational and patterned behavior description interviews: A comparison of their validity, correlates, and perceived fairness. *Human Performance*, *16*(1), 25–47. https://doi.org/10.1207/S15327043HUP1601_2

Dunlop, P. D., Holtrop, D., & Wee, S. (2022). How asynchronous video interviews are used in practice: A study of an australian-based avi vendor. *International Journal of Selection and Assessment*, *30*(3), 448–455. https://doi.org/10.1111/ijsa.12372

Ellis, A. P., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology*, *87*(6), 1200–1208. https://doi.org/10.1037//0021-9010.87.6.1200

Endrass, B., Rehm, M., André, E., & Nakano, Y. I. (2008). Talk is silver, silence is golden: A cross cultural study on the usage of pauses in speech. *Proceedings of the IUI Workshop on Enculturating Conversational Interfaces*.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538. https://doi.org/10.1126/science.1168244

Feiler, A. R., & Powell, D. M. (2016). Behavioral expression of job interview anxiety. *Journal of Business and Psychology*, *31*(1), 155–171. https://doi.org/10.1007/s10869-015-9403-z

Finnerty, A. N., Muralidhar, S., Nguyen, L. S., Pianesi, F., & Gatica-Perez, D. (2016). Stressful first impressions in job interviews. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 325–332. https://doi.org/10.1145/2993148.2993198

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*(6), 709–738. https://doi.org/10.1006/jmla.1995.1032

Gao, G., & Ting-Toomey, S. (1998). *Communicating effectively with the chinese*. Thousand Oaks, CA: Sage Publications.

Gebhard, P., Schneeberger, T., André, E., Baur, T., Damian, I., Mehlmann, G., König, C., & Langer, M. (2019). Serious games for training social skills in job interviews. *IEEE Transactions on Games*, *11*(4), 340–351. https://doi.org/10.1109/TG.2018.2808525

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing re-

port where human-labeled training data comes from? *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain*, 325–336. https://doi.org/10.1145/3351095.3372862

Gifford, R., Ng, C. F., & Wilkinson, M. (1985). Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, *70*(4), 729–736. https://doi.org/10.1037/0021-9010.70.4.729

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.2307/258595

Glassdoor Team. (2015, September 18). *The true cost of a bad hire*. https://www.glassdoor.com/blog/the-true-cost-of-a-bad-hire/

Goffman, E. (1956). *Presentation of self in everyday life*. Social Sciences Research Centre: University of Edinburgh.

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face interaction*. Pantheon Books, New York.

Green, M. C., & Appel, M. (2024). Narrative transportation: How stories shape how we see ourselves and the world. *Advances in Experimental Social Psychology*, *70*(1), 1–82. https://doi.org/10.1016/bs.aesp.2024.03.002

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Griswold, K. R., Phillips, J. M., Kim, M. S., Mondragon, N., Liff, J., & Gully, S. M. (2022). Global differences in applicant reactions to virtual interview synchronicity. *The International Journal of Human Resource Management*, *33*(15), 2991–3018. https://doi.org/10.1080/09585192.2021.1917641

Hartwell, C. J., Johnson, C. D., & Posthuma, R. A. (2019). Are we asking the right questions? predictive validity comparison of four structured interview question types. *Journal of Business Research*, *100*, 122–129. https://doi.org/10.1016/j.jbusres.2019.03.026

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

Heimann, A. L., Ingold, P. V., & Kleinmann, M. (2020). Tell us about your leadership style: A structured interview approach for assessing leadership behavior constructs. *The Leadership Quarterly*, *31*(4), 101364. https://doi.org/10.1016/j.leaqua.2019.101364

Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.-C., & Clavel, C. (2019). Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 573–581. https://doi.org/10.1609/aaai.v33i01.3301573

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323. https://doi.org//10.1037/apl0000695

Hirevue. (2021). Hirevue customers conduct over 1 milion video interviews in just 30 days. https://www.hirevue.com/press-release/hirevue-customers-conduct-over-1-million-video-interviews-in-just-30-days

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5). Sage Publications.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Culture and organizations. software of the mind: Intercultural cooperation and its importance for survival*. NewYork, NY : McGraw Hill. https://doi.org/10.1080/00208825.1980

Hoque, M., Courgeon, M., Martin, J.-C., Mutlu, B., & Picard, R. W. (2013). Mach: My automated conversation coach. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 697–706. https://doi.org/10.1145/2493432.2493502

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*(2), 184–190. https://doi.org/10.1037/0021-9010.79.2.184

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U.-C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment*, *12*(3), 262–273. https://doi.org/10.1111/j.0965-075X.2004.280_1.x

Huffcutt, A. I., & Murphy, S. A. (2023). Structured interviews: Moving beyond mean validity... *Industrial and Organizational Psychology*, *16*(3), 344–348. https://doi.org/10.1017/iop.2023.42

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72–98. https://doi.org/10.1037/0033-2909.96.1.72

Imada, A. S., & Hakel, M. D. (1977). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology*, *62*(3), 295–300. https://doi.org/10.1037/0021-9010.62.3.295

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, *56*(2), 291–303. https://doi.org/10.1016/j.jml.2006.07.011

Jan, J., Alshare, K. A., & Lane, P. L. (2022). Hofstede's cultural dimensions in technology acceptance models: A meta-analysis. *Universal Access in the Information Society*, *23*, 1–25. https://doi.org/10.1007/s10209-022-00930-7

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/0.1007/s12525-021-00475-2

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Sage Publications.

Jaworski, A. (1993). *The power of silence: Social and pragmatic perspectives* (Vol. 1). Sage Publications.

Kacmar, K. M., Delery, J. E., & Ferris, G. R. (1992). Differential effectiveness of applicant impression management tactics on employment interview decisions 1. *Journal of Applied Social Psychology*, *22*(16), 1250–1272. https://doi.org/10.1111/j.1559-1816.1992.tb00949.x

Kessler, R. (2006). *Competency-based interviews: Master the tough new interview style and give them the answers that will win you the job*. Frankilin Lakes : Career Press.

Khan, P., Kader, M. F., Islam, S. R., Rahman, A. B., Kamal, M. S., Toha, M. U., & Kwak, K.-S. (2021). Machine learning and deep learning approaches for brain disease diagnosis: Principles and recent advances. *IEE Access*, *9*, 37622–37655. https://doi.org/10.1109/ACCESS.2021.3062484

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

Kleinlogel, E. P., Schmid Mast, M., Jayagopi, D. B., Shubham, K., & Butera, A. (2023). "The interviewer is a machine!" Investigating the effects of conventional and technol-ogy-mediated interview methods on interviewee reactions and behavior. *International Journal of Selection and Assessment*, *31*(3), 403–419. https://doi.org/10.1111/ijsa.12433

Kohn, L. S., & Dipboye, R. L. (1998). The effects of interview structure on recruiting outcomes. *Journal of Applied Social Psychology*, *28*(9), 821–843. https://doi.org/10.1111/j.1559-1816.1998.tb01733.x

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., van Breda, W., Zhang, T., & de Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance

evaluation using machine learning. *Computers in Human Behavior*, 108128. https://doi.org/10.1016/j.chb.2023.108128

Krajewski, H. T., Goffin, R. D., McCarthy, J. M., Rothstein, M. G., & Johnston, N. (2006). Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future? *Journal of Occupational and Organizational Psychology*, *79*(3), 411–432. https://doi.org/10.1348/096317905X68790

Krause, R. J., & Rucker, D. D. (2020). Strategic storytelling: When narratives help versus hurt the persuasive power of facts. *Personality and Social Psychology Bulletin*, *46*(2), 216–227. https://doi.org/10.1177/0146167219853

Langer, M., & König, C. J. (2018). Introducing and testing the creepiness of situation scale (cross). *Frontiers in Psychology*, *9:2220*. https://doi.org/10.3389/fpsyg.2018.02220

Langer, M., König, C. J., Gebhard, P., & André, E. (2016). Dear computer, teach me manners: Testing virtual employment interview training. *International Journal of Selection and Assessment*, *24*(4), 312–323. https://doi.org/10.1111/ijsa.12150

Langer, M., König, C. J., & Hemsing, V. (2020). Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions. *Journal of Managerial Psychology*, *35*(4), 271–284. https://doi.org/10.1108/JMP-03-2019-0156

Langer, M., König, C. J., & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. *International Journal of Selection and Assessment*, *25*(4), 371–382. https://doi.org/10.1111/ijsa.12191

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217–234. https://doi.org/10.1111/ijsa.12246

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology*, *65*(4), 422–427. https://doi.org/10.1037/0021-9010.65.4.422

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, *107*(1), 34–77. https://doi.org/10.1037/0033-2909.107.1.34

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

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

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer. https://doi.org/10.1007/978-3-319-98131-4_9

Lievens, F., & De Paepe, A. (2004). An empirical investigation of interviewer-related factors that discourage the use of high structure interviews. *Journal of Organizational Behavior*, *25*(1), 29–46. https://doi.org/10.1002/job.246

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology*, *109*(6), 921–948. https://doi.org/10.1037/apl0001173

Lin-Stephens, S., Manuguerra, M., Tsai, P.-J., & Athanasou, J. A. (2022). Stories of employability: Improving interview narratives with image-supported past-behaviour storytelling training. *Education + Training*, *64*(5), 577–597. https://doi.org/10.1108/ET-08-2021-0320

Liu, R. W., Lapinski, M. K., Kerr, J. M., Zhao, J., Bum, T., & Lu, Z. (2022). Culture and social norms: Development and application of a model for culturally contextualized communication measurement (MC3M). *Frontiers in Communication*, *6*, 770513. https://doi.org/10.3389/fcomm.2021.770513

Lukacik, E.-R., & Bourdage, J. S. (2024). Does design matter? The (limited) effects of six asynchronous video interview design features on impression management, reactions, and evaluations. *International Journal of Selection and Assessment*, *33*(1), 1–18. https://doi.org/10.1111/ijsa.12511

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), e12511. https://doi.org/10.1016/j.hrmr.2020.100789

Mandelbaum, J. (2013). Storytelling in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 492–507). Chichester, UK: John Wiley & Sons.

Manroop, L., Boekhorst, J. A., & Harrison, J. A. (2013). The influence of cross-cultural differences on job interview selection decisions. *The International Journal of Human Resource Management*, *24*(18), 3512–3533. https://doi.org/10.1080/09585192.2013.777675

Maurer, T. J., & Solamon, J. M. (2006). The science and practice of a structured employment interview coaching program. *Personnel Psychology*, *59*(2), 433–456. https://doi.org/10.1111/j.1744-6570.2006.00797.x

McLaren, S. (2019, March 4). *The tactic this expert uses to assess soft skills*. LinkedIn. https://www.linkedin.com/business/talent/blog/talent-acquisition/tactic-this-expert-uses-to-assess-soft-skills

Melchers, K. G., Petrig, A., Basch, J. M., & Sauer, J. (2021). A comparison of conventional and technology-mediated selection interviews with regard to interviewees' performance, perceptions, strain, and anxiety. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.603632

Nabi, R. L. (2015). Emotional flow in persuasive health messages. *Health Communication*, *30*(2), 114–124. https://doi.org/10.1080/10410236.2014.974129

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, *1*, 1–6. https://doi.org/10.1109/FG.2015.7163127

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, *16*(4), 1018–1031. https://doi.org/10.1109/TMM.2014.2307169

Orji, K., & Bangerter., A. (2024). *Tick tock, tick tock, time's up! how response time limits in asynchronous video interviews shape interview anxiety, impression management, and performance* [Manuscript in preparation]. Institute of Work and Organizational Psychology, University of Neuchâtel.

Orji, K., Bangerter., A., Germanier, E., Renier, L. A., Mast, S., M., M., He, & Garner, P. N. (2024). *Extended responses in asynchronous video interviews: Investigating frequency, content, and interview outcomes* [Manuscript in preparation]. Institute of Work and Organizational Psychology, University of Neuchâtel.

Peck, J. A., & Levashina, J. (2017). Impression management and interview and job performance ratings: A meta-analysis of research design with tactics in mind. *Frontiers in Psychology*, *8*, 201. https://doi.org/10.3389/fpsyg.2017.00201

Piolat, A., Booth, R., Chung, C. K., Davids, M., & Pennebaker, J. (2011). La version française du dictionnaire pour le LIWC: Modalités de construction et exemples d'utilisation. *Psychologie Française*, *56*(3), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002

Posthuma, R. A., Levashina, J., Lievens, F., Schollaert, E., Tsai, W.-C., Wagstaff, M. F., & Campion, M. A. (2014). Comparing employment interviews in latin america

with other countries. *Journal of Business Research*, *67*(5), 943–951. https://doi.org/10.1016/j.jbusres.2013.07.014

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

Potosky, D. (2008). A conceptual framework for the role of the administration medium in the personnel assessment process. *Academy of Management Review*, *33*(3), 629–648. https://doi.org/10.5465/amr.2008.32465704

Proost, K., Schreurs, B., De Witte, K., & Derous, E. (2010). Ingratiation and self-promotion in the selection interview: The effects of using single tactics or a combination of tactics on interviewer judgments. *Journal of Applied Social Psychology*, *40*(9), 2155–2169. https://doi.org/10.1111/j.1559-1816.2010.00654.x

Rahman, W., Mahbub, S., Salekin, A., Hasan, M. K., & Hoque, E. (2021). Hirepreter: A framework for providing fine-grained interpretation for automated job interview analysis. *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 1–5. https://doi.org/10.1109/ACIIW52867.2021.9666201

Ralston, S. M., Kirkwood, W. G., & Burant, P. A. (2003). Helping interviewees tell their stories. *Business Communication Quarterly*, *66*(3), 8–22. https://doi.org/10.1177/108056990306600303

Rasmussen, K. G. (1984). Nonverbal behavior, verbal behavior, resumé credentials, and selection interview outcomes. *Journal of Applied Psychology*, *69*(4), 551–556. https://doi.org/10.1037/0021-9010.69.4.551

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, *5*, 31. https://doi.org/10.1140/epjds/s13688-016-0093-1

Riggio, R. E., & Throckmorton, B. (1988). The relative effects of verbal and nonverbal behavior, appearance, and social skills on evaluations made in hiring interviews. *Journal of Applied Social Psychology*, *18*(4), 331–348. https://doi.org/0.1111/j.1559-1816.1988.tb00020.x

Rizi, M. S., & Roulin, N. (2024). Does media richness influence job applicants' experience in asynchronous video interviews? Examining social presence, impression management, anxiety, and performance. *International Journal of Selection and Assessment*, *32*(1), 54–68. https://doi.org/10.1111/ijsa.12448

Rosenfeld, P. (1997). Impression management, fairness, and the employment interview. *Journal of Business Ethics*, *16*(8), 801–808. https://doi.org/10.1023/A:1017972627516

Roulin, N., Bangerter, A., & Wüthrich, U. (2012). *Réussir l'entretien d'embauche comportemental: La méthode pour identifier et sélectionner les futurs employés performants*. Bruxelles: De Boeck Professionals.

Roulin, N., Pham, L. K. A., & Bourdage, J. S. (2023). Ready? Camera rolling... action! Examining interviewee training and practice opportunities in asynchronous video interviews. *Journal of Vocational Behavior*, *145*, 103912. https://doi.org/10.1016/j.jvb.2023.103912

Roulin, N., Wong, O., Langer, M., & Bourdage, J. S. (2023). Is more always better? How preparation time and re-recording opportunities impact fairness, anxiety, impression management, and performance in asynchronous video interviews. *European Journal of Work and Organizational Psychology*, *32*(3), 333–345. https://doi.org/10.1080/1359432X.2022.2156862

Rupasinghe, A. T., Gunawardena, N. L., Shujan, S., & Atukorale, D. (2016). Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis. *2016 16th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 288–295. https://doi.org/10.1109/ICTER.2016.7829933

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040. https://ink.library.smu.edu.sg/lkcsb_research/6894

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

Sandal, G. M., van de Vijver, F., Bye, H. H., Sam, D. L., Amponsah, B., Cakar, N., Franke, G. H., Ismail, R., Kjellsen, K., & Kosic, A. (2014). Intended self-presentation tactics in job interviews: A 10-country study. *Journal of Cross-Cultural Psychology*, *45*(6), 939–958. https://doi.org/10.1177/00220221145323

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations* (Vol. 526). Monterey, CA: Brooks/Cole.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262. https://doi.org/10.1037/0033-2909.124.2.262

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology*, *77*(5), 629. https://doi.org/10.1037/0021-9010.77.5.629

Schneider, D. J. (1981). Tactical self-presentations: Toward a broader conception. In J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp. 23–40). New York : Academic Press.

Schneider, L., Powell, D. M., & Bonaccio, S. (2019). Does interview anxiety predict job performance and does it influence the predictive validity of interviews? *International Journal of Selection and Assessment*, *27*(4), 328–336. https://doi.org/10.1111/ijsa.12263

Sekine, K., & Kita, S. (2017). The listener automatically uses spatial story representations from the speaker's cohesive gestures when processing subsequent sentences without gestures. *Acta Psychologica*, *179*, 89–95. https://doi.org/10.1016/j.actpsy.2017.07.009

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. Bath: John Wiley & Sons.

Silvester, J. (1997). Spoken attributions and candidate success in graduate recruitment interviews. *Journal of Occupational and Organizational Psychology*, *70*(1), 61–73. https://doi.org/10.1111/j.2044-8325.1997.tb00631.x

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, *9*(6), 838–846. https://doi.org/10.1037/a0017810

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*(2), 149–155. https://doi.org/10.1037/h0047060

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Smither, J. W., Reilly, R. R., Millsap, R. E., AT&T, K. P., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, *46*(1), 49–76. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, *80*(5), 587–606. https://doi.org/10.1037/0021-9010.80.5.587

Stone, D. L., Deadrick, D. L., Lukaszewski, K. M., & Johnson, R. (2015). The influence of technology on the future of human resource management. *Human Resource Management Review*, *25*(2), 216–231. https://doi.org/10.1016/j.hrmr.2015.01.002

Suen, H.-Y., Hung, K.-E., & Lin, C.-L. (2019). Tensorflow-based automatic personality recognition used in asynchronous video interviews. *IEEE Access*, *7*, 61018–61023. https://doi.org/10.1109/ACCESS.2019.2902863

Sy, T., Horton, C., & Riggio, R. (2018). Charismatic leadership: Eliciting and channeling follower emotions. *The Leadership Quarterly*, *29*(1), 58–69. https://doi.org/10.1016/j.leaqua.2017.12.008

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology*, *75*(3), 277–294. https://doi.org/10.1348/096317902320369712

Tessler, R., & Sushelsky, L. (1978). Effects of eye contact and social status on the perception of a job applicant in an employment interviewing situation. *Journal of Vocational Behavior*, *13*(3), 338–347. https://doi.org/10.1016/0001-8791(78)90060-X

The Culture Factor Group. (2024). *Country Comparison Tool*. https://www.theculturefactor.com/country-comparison-tool?countries=india%2Cswitzerland

Triandis, H. C. (1982). Review of culture's consequences: International differences in work-related values. *Human Organization*, *41*(1), 86–90.

UNIKO Media Group. (2024, July 25). *The cost of a bad hire: Quantifying the impact of poor hiring decisions*. Plexus Global. https://plexusglobalinc.com/the-cost-of-a-bad-hire-quantifying-the-impact-of-poor-hiring-decisions/

Uono, S., & Hietanen, J. K. (2015). Eye contact perception in the west and east: A cross-cultural study. *Plos one*, *10*(2), e0118094. https://doi.org/10.1371/journal.pone.0118094

Van der Zee, K. I., Bakker, A. B., & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology*, *87*(1), 176–184. https://doi.org/10.1037//0021-9010.87.1.176

Van Iddekinge, C. H., McFarland, L. A., & Raymark, P. H. (2007). Antecedents of impression management use and effectiveness in a structured interview. *Journal of Management*, *33*(5), 752–773. https://doi.org/10.1177/014920630730556

Van Iddekinge, C. H., Raymark, P. H., Roth, P. L., & Payne, H. S. (2006). Comparing the psychometric characteristics of ratings of face-to-face and videotaped structured interviews. *International Journal of Selection and Assessment*, *14*(4), 347–359. https://doi.org/10.1111/j.1468-2389.2006.00356.x

Van Laer, T., De Ruyter, K., Visconti, L. M., & Wetzels, M. (2014). The extended transportation-imagery model: A meta-analysis of the antecedents and consequences of consumers' narrative transportation. *Journal of Consumer Research*, *40*(5), 797–817. https://doi.org/10.1086/673383

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vijay, R. S., Shubham, K., Renier, L. A., Kleinlogel, E. P., Mast, M. S., & Jayagopi, D. B. (2021). An opportunity to investigate the role of specific nonverbal cues and first impression in interviews using deepfake based controlled video generation. *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 148–152. https://doi.org/10.1145/3461615.3485397

Wasielewski, P. L. (1985). The emotional basis of charisma. *Symbolic Interaction*, *8*(2), 207–222.

Wilhelmy, A., Kleinmann, M., König, C. J., Melchers, K. G., & Truxillo, D. M. (2016). How and why do interviewers try to make impressions on applicants? A qualitative study. *Journal of Applied Psychology*, *101*(3), 313–332. https://doi.org/10.1037/apl0000046

Winkler, J. R., Mengelkamp, C., & Appel, M. (2022). Real-time responses to stories: Linking valence shifts to post-exposure emotional flow and transportation. *Communication Research Reports*, *39*(5), 237–247. https://doi.org/10.1080/08824096.2022.2119380

Appendices

I Appendix 1: Study 1

Germanier, E., Bangerter, A., Orji, K., Schmid Mast, M., Renier, L. A., He, M., & Garner, P. N. (2024). *The Effects of Applicants' Emotional Framing in Narrative Interview Responses on Interview Outcomes* [Manuscript in preparation]. Institute of Work and Organizational Psychology, University of Neuchâtel.

Status: The manuscript is currently in preparation. The version in this thesis is the latest one.

Effect of Emotional Framing in Narrative Responses to Past-Behavior Questions on Interview Outcomes

Elisabeth Germanier[1] , Adrian Bangerter[1] , Koralie Orji[1] , Laetitia A. Renier[2] , Marianne Schmid Mast[2], Mutian He[3], Philip N. Garner[3]

[1] Institute of Work and Organizational Psychology, University of Neuchâtel, Switzerland

[2] Department of Organizational Behavior, University of Lausanne, Switzerland

[3] Idiap Research Institute, Martigny, Switzerland

Author Note

The authors report no conflict of interest.

This work was supported by the Swiss National Science Foundation (SNSF) under Grant 10521C_197479.

Correspondence concerning this article should be addressed to Elisabeth Germanier, Institute of Work and Organizational Psychology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. E-Mail: elisabeth.germanier@unine.ch.

Abstract

Storytelling responses to past-behavior questions are effective for applicants to manage impressions. However, the factors contributing to highly-rated responses remain underexplored. This study explores how emotional framing in narrative responses influences recruiters' perceptions and evaluations of applicants. Using data from mock in-person interviews (n = 128) that included three past-behavior questions, we extracted both positive and negative emotions from transcribed narrative responses and analyzed how their emotional trajectories affect recruiters' perceptions of competence, performance, persuasiveness, engagement, and their emotional engagement with the responses. Results indicated that increasing positive emotions throughout the response led to higher perceptions of persuasiveness, while increasing negative emotions resulted in lower ratings for competence and performance. However, raters' emotional engagement with the response was not predicted by either emotional trajectory.

.

Effects of Emotional Framing in Narrative Interview Responses on Interview Outcomes

Introduction

Behavioral interviews are recommended for hiring organizations as one of the most valid tools to predict applicants' future job performance (Taylor & Small, 2002). These interviews typically include questions about applicants' past behaviors (e.g., "Can you describe a situation where you had to deal with an angry client?"). Past-behavior questions help recruiters assess applicants' characteristics (Campion et al., 1997) by inviting them to describe past work experiences. Ideally, applicants' responses should be stories illustrating how they solved a problematic situation (Bangerter et al., 2014; Brosy et al., 2020; Stevens & Kristof, 1995). However, storytelling in such stressful situations is a challenging task (Brosy et al., 2020). Some applicants can provide well-crafted stories, while others produce sub-optimal narratives describing generic situations (i.e., pseudo-stories; Bangerter et al., 2014). Hence, there is great variety in the production of narrative responses.

To be effective, narrative responses should provide detailed accounts of situations and steps to achieve goals clearly and coherently (Kessler, 2006; Ralston et al., 2003). Narratives are structured by their opening (staging), progress toward the end (plot progression), and development of cognitive tension (Boyd et al., 2020). Still, their effectiveness may also hinge on the evolution of emotions they convey (i.e., emotional trajectories). Narratives can diverge in how positive and negative emotions are introduced and fluctuate throughout the narrative (Reagan et al., 2016). Indeed, a story may open with negative staging and follow a plot progression toward a heroic ending, while another may open with positive staging and follow a plot progression toward a negative ending. These different emotional trajectories may create different cognitive tension and will likely generate different effects on the audience. In the context of selection interviews, a best-rated narrative response may align plot progression and

cognitive tension with an emotional trajectory characterized by increasing positive and decreasing negative emotions. This trajectory would signal successful progress toward a goal by overcoming difficulties. However, little is known about how such emotional trajectories may manifest in narrative responses to past-behavior questions in this formalized type of interaction.

Narrative responses with effective emotional framing will likely impact recruiters' evaluations of applicants. First, narratives allow applicants to attribute characteristics to the protagonist (i.e., themselves; Silvester, 1997). As they fully engage with the problem and successfully navigate challenges in their narratives, recruiters may view applicants who share narratives as more competent and engaged in their jobs. Second, when combined with appropriate emotional framing, narratives can hold a greater persuasive power. Indeed, the trajectory of emotions within narratives influences the audience's reactions (Del Vecchio et al., 2021; Reagan et al., 2016) and behaviors (Alam & So, 2020; Carrera et al., 2008). Consequently, applicants who provide well-framed narrative responses may receive higher evaluations. However, applicants' communication skills may contribute to variability in structured interview validity (Huffcutt & Murphy, 2023). It is thus necessary to understand to what extent emotional framing in narrative responses affects recruiters' evaluation, an underexplored research topic.

This study examines the trajectories of positive and negative emotions in narrative responses to past-behavior questions in mock behavioral interviews. We reanalyzed data from an experimental study in which participants completed face-to-face mock interviews that included past-behavior questions. Positive and negative emotion words were extracted from transcribed narrative responses to past-behavior questions using the Linguistic Inquiry and Word Count (LIWC) software (Boyd et al., 2022; Piolat et al., 2011). Narrative responses were also evaluated based on the video recordings for different criteria (e.g., competence,

performance). This approach enables us to observe how positive and negative emotional trajectories are articulated in interview narrative responses and investigate their impact on evaluations of participants' and raters' emotional engagement with the responses.

This research contributes to personnel selection, both research and practice. First, it expands our knowledge of applicants' narrative responses during job interviews, addressing two key issues. On the one hand, it clarifies what makes a narrative response effective (Bangerter et al., 2014). On the other hand, it explores how applicants' communication, particularly their response framing, may influence recruiters' evaluations. Therefore, it offers an initial answer regarding how applicants' communication relates to variability in structured interview validity (Huffcutt & Murphy, 2023). Second, this research may bring attention to subtle sources of biases in evaluations within organizations. Understanding how response framing might interfere with recruiters' evaluations may help hiring organizations adopt strategies to reduce bias, such as objectively focusing on described behaviors.

Narrative Responses to Past-Behavior Questions

Past-behavior questions prompt applicants to share experiences where they successfully tackled work-related challenges. They are advised to respond by describing the situation, their roles and actions, and the results achieved (Kessler, 2006). Recruiters then evaluate applicants' competencies based on the described behaviors, assuming these behaviors are representative of those competencies (Janz, 1989). To reduce bias, recruiters are advised to use behaviorally anchored rating scales (BARS) for their evaluations (Roulin et al., 2012; Smith & Kendall, 1963).

To answer past-behavior questions, storytelling responses are particularly interesting for applicants because they help foster more vivid and memorable impressions (Stevens & Kristof, 1995), and prove more persuasive than a simple list of facts, particularly when those facts are weak (Krause & Rucker, 2020). Moreover, applicants, as the central figures in their narratives,

can indirectly highlight their positive characteristics and values (Silvester, 1997). By immersing themselves and their audience in the challenges of their stories and demonstrating how they overcame obstacles, applicants who share their stories may be perceived as more competent and engaged in their work.

Telling stories on the fly during job interviews is challenging and cognitively demanding. Applicants must recall a suitable past event, organize the response, and deliver it adroitly within a short timeframe (Brosy et al., 2020). In practice, many struggle to produce fully developed stories without recruiters' assistance (Bangerter et al., 2014; Brosy et al., 2020). Instead, they may turn to pseudo-stories or decontextualized assertions about themselves (e.g., self-description; Bangerter et al., 2014). Yet only narrative responses (i.e., stories and pseudo-stories) were associated with higher hirability ratings (Bangerter et al., 2014). Still, 58.3 % of narrative responses reflect day-to-day typical performance rather than maximal performance, which can be improved by question formulations priming applicants to recall their best performance (Huffcutt et al., 2024). Nevertheless, we still lack a clear understanding of what constitutes an effective narrative response – specifically, what makes one narrative performance better than another and how this influences interview outcomes.

Emotional Framing in Narratives

Effective narrative responses may be characterized by applicants' ability to manipulate the framing of their narratives effectively. Indeed, the same narrative content (*what is said*) can be presented in different ways (*how it is said*), thus affecting recruiters differently (Imada & Hakel, 1977). For instance, one applicant might recount a situation where an angry person came to the front office and that they found a solution that seemed to satisfy the client, while another better equipped for impactful communication may emphasize the severity of the initial situation (e.g., yelling in front of other clients) and highlight their heroic resolution (e.g., the client ending up effusively thanking the staff and leaving with a smile).

Narratives follow common structural and emotional patterns. Many narratives (e.g., fairy tales, film scripts, newspaper articles, TED Talks) are structured similarly: They typically open with an introduction of the situation (*staging*) followed by the development of the plot (*plot progression*) toward the narrative's resolution, while *cognitive tension* builds in parallel, reaching its peak before the story ends (Boyd et al., 2020). Narratives also share patterns in how emotions unfold, thus following core emotional trajectories (see Reagan et al., 2016). One such trajectory is the progressive rise of positive emotions, with decreasing negative emotions. Some emotional trajectories are linked to narrative success with the audience, though this success may vary across formats (e.g., written vs audio-visual narrative; Del Vecchio et al., 2021; Reagan et al., 2016).

Emotional trajectories contribute to shaping the audience's reactions. For instance, in health communication interventions, narratives that shift from negative to positive emotions (or vice versa) induce more behavioral change than narratives revolving around a single constant emotion (Alam & So, 2020; Carrera et al., 2008; Nabi, 2015). Hence, a narrative showing a decrease in negative emotions alongside an increase in positive emotions is more likely to induce reactions in the audience than one lacking emotional shift. Similarly, in organizational contexts, charismatic leaders harness emotional trajectories to reinforce their message and create emotional shifts in the audience. By first invoking the audience's emotions, then revoking, and finally reframing them (Wasielewski, 1985), they skillfully steer the audience toward the desired change (Sy et al., 2018).

The persuasive power of certain emotional trajectories in narratives can facilitate the narrative transportation of the audience. When individuals engage with a story, they may become absorbed in the evocative imagery conveyed by the protagonist and transported by the emotions in the narrative (Winkler et al., 2022). Their attention, emotions, and imagination become intricately tied to the narrative, fostering an emotional connection (Green & Appel,

2024; Green & Brock, 2000). This process of narrative transportation lessens defensive information processing, making individuals more receptive to the narrative's message (Green & Brock, 2000; Krause & Rucker, 2020; Van Laer et al., 2014). Emotional framing can enhance the emotional connection with the narratives, reinforcing audience transportation and shaping their reactions. As a result, positive and negative emotional trajectories within narratives may strengthen the narrative's persuasive power and influence the audience's beliefs and behaviors.

This study

This study investigates what constitutes an effective narrative response to past-behavior questions in structured interviews through the lens of emotional framing. Since narratives all share one of the core emotional trajectories (Del Vecchio et al., 2021; Reagan et al., 2016), which can influence the audience's reactions and behaviors (Alam & So, 2020; Carrera et al., 2008; Sy et al., 2018), applicants' narrative responses should also incorporate emotional trajectories likely to affect recruiters' evaluations. Moreover, some applicants may adjust their narrative responses using efficient emotional framing to enhance their persuasive effect by anticipating what will resonate most. When framed effectively, these emotional trajectories may enhance the response's impact and thus positively affect recruiters' evaluations of their performance and engagement. In particular, best-rated narrative responses would structure their emotional trajectories to suggest successful progress toward a goal by overcoming difficulties. Such framing would correspond to an emotional trajectory of increasing positive and decreasing negative emotions. To our knowledge, there has been no research on the emotional trajectories involved in narrative responses to past-behavior questions, nor how these might impact recruiters' evaluations of applicants or their emotional engagement with the responses. Therefore, we ask how positive and negative emotions fluctuate over the course of narrative responses to past-behavior questions (Research Question 1; RQ1) and how positive and

negative emotional trajectories within narrative responses affect recruiters' evaluations and emotional engagement with the response (Research Question 2; RQ2).

The present study used part of Germanier et al.'s (2024) data. In their experiment, participants completed a work sample (i.e., delivering bad news) and a mock interview one week later in the face-to-face condition with a mock recruiter or in the asynchronous video interview condition, videorecording their responses through the computer. Those interviews included three past-behavior questions, with the last question focusing on the previous week's work sample. For this study, we focused on face-to-face interviews only due to fundamental differences between the two interview settings, preventing us from comparisons. In traditional interview settings, the recruiter plays a crucial role in co-creating applicants' narratives by providing real-time feedback and using backchannels (Mandelbaum, 2013). However, this co-creation is absent in asynchronous video interviews, where applicants must respond using their computer without the benefit of live interaction (Lukacik et al., 2022). Comparing narrative responses from these two interview settings would lead to skewed interpretations, as one involves co-constructed narratives and the other does not. Moreover, asynchronous video interview responses differ significantly from those in face-to-face settings, both in content and structure (Germanier et al., 2024; Orji et al., 2024).

To address RQ1 and RQ2, we use the video recordings of the responses to measure raters' evaluations and emotional engagement and the transcribed narrative responses to measure positive and negative emotions. To observe the emotional progression over the course of a narrative response, we divided each response into five equal parts based on the number of words (i.e., quintiles), following Boyd et al.'s (2020) method. This segmentation allows us to compare the proportions of positive and negative emotions across responses while controlling for variation in the length of responses. At the same time, it offers a reasonable trade-off between short interview narrative responses and the observation of progression throughout the

response. Interview narrative responses are shorter than traditional written narratives that typically contain a much greater number of words. Dividing interview narrative responses into smaller segments might result in segments with too few words for meaningful analysis. In each quintile narrative response, we measured positive and negative emotions, extracting the percentage of words with positive and negative emotional valence using LIWC (Boyd et al., 2022; Piolat et al., 2011). This approach helps us observe the tendency for positive and negative emotional trajectories in narrative responses and examine their effect on the rater's evaluations of performance, engagement, persuasiveness, and the rater's emotional engagement. We also examine their impact on competence as measured by BARS, which is regarded as a more objective measure focusing on observable behaviors by anchoring evaluations in specific examples of behavior thus more robust to biases (Roulin et al., 2012; Smith & Kendall, 1963). As such, we can compare the effect of emotional response framing on "objective" versus more general "subjective" measures.

Method

Sample

Participants were recruited through a pool of participants from a Swiss university's online recruitment system. A hundred twenty-eight French-speaking students completed a face-to-face mock interview (59.38% female, *M*age = 21.61, *SD* = 2.90). At the time of the experiment, 52.34% of the participants had completed a high school degree, 34.38% had obtained their bachelor's degree, 12.5% held a master's degree, and one participant reported having another type of degree. On average, participants had past 1.90 experiences of job interviews (*SD* = 2.75), and 78.91% had previous work experience.

Since we want to explore emotional trajectory in narrative responses, we excluded four responses (from four participants) that contained no narrative utterance. Plus, we lack one

participant's response because of recording failure (final N = 123 participants, one of which provided three responses, and 5 who provided two responses only).

Procedure

Participants read and signed a consent form about one week before the mock interview (at the beginning of the work sample session; see Germanier et al., 2024). When participants arrived at the lab for the face-to-face mock interview, the experimenter briefly informed them about the job interview for a manager position in a company selling computers. The recruiter was played by an I/O psychology graduate student trained to follow the job interview script. The recruiter welcomed the participant and invited them to sit at the table, thanked them for coming to the interview, explained the interview procedure, and asked them four questions. After the interview, participants filled out the Honest Interview Impression Management questionnaire (Bourdage et al., 2018; data not reported here).

The first question served as a warm-up question inviting participants to present themselves. The three other questions were past behavior questions targeting managerial competencies: multitasking ("Can you describe a situation where you were responsible for completing several tasks in a short time?"), team management ("Can you describe a situation where one of your colleagues did not complete the tasks requested within the time limit? How did you deal with this situation?") and delivering bad news ("Can you tell us about a recent situation in which you had to give bad news to someone?"). All participants answered the warm-up questions first and ended the interview with the bad news delivery questions, and the order of the multitasking and team management questions were counterbalanced. Throughout the interview, the recruiter could produce backchannels (e.g., "mhm*"*) or use words like "indeed" to show active listening.

Participants were audio and video recorded during the interview with a microphone and front and side cameras. The job interviews were transcribed word for word based on the audio recordings.

Measures

Competence

Competence was assessed by two trained psychology undergraduates based on side video recordings of responses to past-behavior questions using self-developed BARS with definitions and examples of behaviors (Smith & Kendall, 1963). For each response, the raters separately watched the video recording and evaluated the response from 1 (=*not at all competent*) to 5 (=*fully competent*). We averaged the two raters' scores to get a single measure for each question. Interrater reliability was computed as the correlation between both ratings of responses (r = .83, $p < .001$).

Performance

Performance was assessed by one trained psychology undergraduate based on side video recordings of responses to past-behavior questions. For each response, the rater coded for perceived competence (i.e., the extent to which participants looked competent), clarity (i.e., the extent to which the answer was clear), and relevance (i.e., the extent to which the answer was relevant to the question), on scales from 1 (=*poor*) to 5 (=*excellent*). To get a single measure of performance for each participant's response, we averaged the three scores (α = .91). Interrater reliability was established using a second independent rater double-coding 24 interviews (r = .85, $p < .001$).

Engagement

Engagement was assessed by one trained psychology undergraduate based on side video recordings of responses to past-behavior questions. For each response, the rater evaluated the extent to which the participant looked motivated on a scale from 1 (=*poor*) to 5 (=*excellent*).

Interrater reliability was established using a second independent rater double-coding 24 interviews ($r = .82$, $p < .001$).

Persuasiveness

Persuasiveness was assessed by one trained psychology undergraduate based on side video recordings of responses to past-behavior questions. For each response, the rater evaluated the extent to which the participant was persuasive on a scale from 1 (=*poor*) to 5 (=*excellent*). Interrater reliability was established using a second independent rater double-coding 24 interviews ($r = .81$, $p < .001$).

Rater's Emotional Engagement

The rater's emotional engagement was assessed by one trained psychology undergraduate based on side video recordings of responses to past-behavior questions. For each response, the rater evaluated to which extent they felt emotionally engaged while listening to a response on a scale from 1 (=*not at all*) to 5 (=*completely*). Interrater reliability was established using a second independent rater, double-coding 24 interviews ($r = .88$, $p < .001$).

Positive and Negative Emotions

The emotion extraction from the participants' responses to past behavior questions was done in three steps. First, we cleaned the transcribed narrative responses to past-behavior questions by removing filler words (e.g., "um", "huh") and annotations into brackets (e.g., laughter, coughing). Second, we divided each response into quintiles (i.e., five equal text sizes in the number of words; Boyd et al., 2022). Third, we extracted the percentage of positive and negative emotion words for each quintile using the 2022 LIWC software (Boyd et al., 2022) with the 2011 LIWC French dictionary (Piolat et al., 2011).

To assess the positive and negative emotional trajectories for each narrative response, we conducted an OLS regression using the quintile as the predictor for the percentage of words expressing positive emotions. This procedure was also applied to negative emotions. We then

used the slopes for positive and negative emotions derived from the model to operationalize the emotional trajectories in each narrative response.

When inspecting the slopes, we noticed that four responses had zero slopes for both emotions, indicating that there were no emotional words in these narrative responses. Since we are interested in the trajectories of emotions and their effects, we excluded those four responses from the present analyses. Thus, we have $N = 121$ participants with three narrative responses, $N = 5$ participants with two narrative responses, and $N = 2$ participants with one narrative response.

Data Analyses

Our data involves participants responding to the same three interview questions, each with different content. The past-behavior questions elicited narratives of emotionally loaded situations that differed (e.g., multitasking versus delivering bad news). Therefore, we employ analytical methods that account for random effects arising from individual and question-related differences. These approaches enable us to consider all narrative responses rather than just the means by the participants and to capture both within-person and between-person variability.

We provide between-subject and within-subject correlations for the study's main variables. To answer RQ1 (description of positive and negative emotional trajectories throughout the response), we provide repeated measures correlations that adjust for inter-individual variability, thus removing the measured variance between participants (Bakdash & Marusich, 2017).

To address RQ2 (the effect of positive and negative emotional trajectories on evaluations), we use linear mixed-effect models with the crossed random effects of individuals and the questions (Baayen et al., 2008). In building our models, we model the fixed effects of the positive and negative emotional trajectories (Slope Positive Emotions, Slope Negative Emotions, Slope Positive Emotions X Slope Negative Emotions; fixed effects). We also controlled for the total amount of positive emotions (i.e., Sum Positive Emotions) and negative

emotions (Sum Negative Emotions; fixed effects). As for the random effects, we started with the maximal random effect structure required by the design (Barr et al., 2013) with random intercepts and slopes for participants and questions. However, that level of random structure was too complex for the model to provide estimates and failed to converge. Therefore, we finally include random intercepts for participants and questions only. All continuous independent variables were centered at their grand mean.

Results

The means and standard deviations of study variables by question are presented in Table 1. The between-subject and within-subject correlations for the main study variables are presented in Table 2. The analyses used R 4.4.1 (R Core Team, 2024) and the following R packages: tidyverse (Wickham et al., 2019) for data manipulation and visualization, psych (Revelle, 2024) for computation of between-subject and within-subject correlations, rmcorr (Bakdash & Marusich, 2024) for computation of repeated-measures correlations, and lme4 (Bates et al., 2014) for mixed linear models with crossed random effects.

Table 1

Descriptive Statistics of Study Variables by Question

	Multitasking			**Team Management**			**Delivering Bad News**		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Competence	123	2.73	1.11	124	2.63	1.18	125	3.02	1.05
Performance	116	2.80	0.79	117	3.05	0.85	118	3.81	0.70
Engagement	116	3.41	0.78	117	3.44	0.85	118	3.64	0.77
Persuasiveness	116	2.90	0.81	117	2.99	0.86	118	3.03	0.77
Rater's Emotional Engagement	116	1.28	0.57	117	1.34	0.60	118	2.11	0.85
Sum Positive Emotions	124	15.68	8.75	125	14.57	8.23	126	17.44	7.46
Sum Negative Emotions	124	6.60	5.51	125	7.96	5.96	126	12.30	7.50
Slope Positive Emotions	124	0.57	0.94	125	0.44	1.19	126	0.52	1.02
Slope Negative Emotions	124	0.01	0.62	125	0.08	0.65	126	-0.11	0.73

Table 2

Between-Subject and Within-Subject Correlation Matrixes

		1	2	3	4	5	6	7	8	9
1	Competence		*.37***	*.25***	*.29***	*-.03*	*.06*	*-.03*	*.14**	*-.18****
2	Performance	.42***		*.40***	*.31***	*.38***	*.08*	*.29***	*-.02*	*-.23****
3	Engagement	.29***	.41***		*.26***	*.31***	*-.05*	*.13**	*.05*	*-.15****
4	Persuasiveness	.28***	.43***	.57***		*.11**	*-.05*	*.11**	*.12**	*-0.08*
5	Rater's Emotional Engagement	.14	.10	.46***	.16		*.09*	*.33****	*.09*	*-.16****
6	Sum Positive Emotions	.20*	-.05	-.23*	-.11	-.04		*-.01*	*.14**	*-.16****
7	Sum Negative Emotions	-.17*	-.07	.03	-.06	.17	-.12		*-.10*	*-.08*
8	Slope Positive Emotions	.03	-.08	-.23*	.10	-.09	.24***	-.12		*-0.04*
9	Slope Negative Emotions	-.03	.06	.02	.04	.06	.05	.24***	.00	

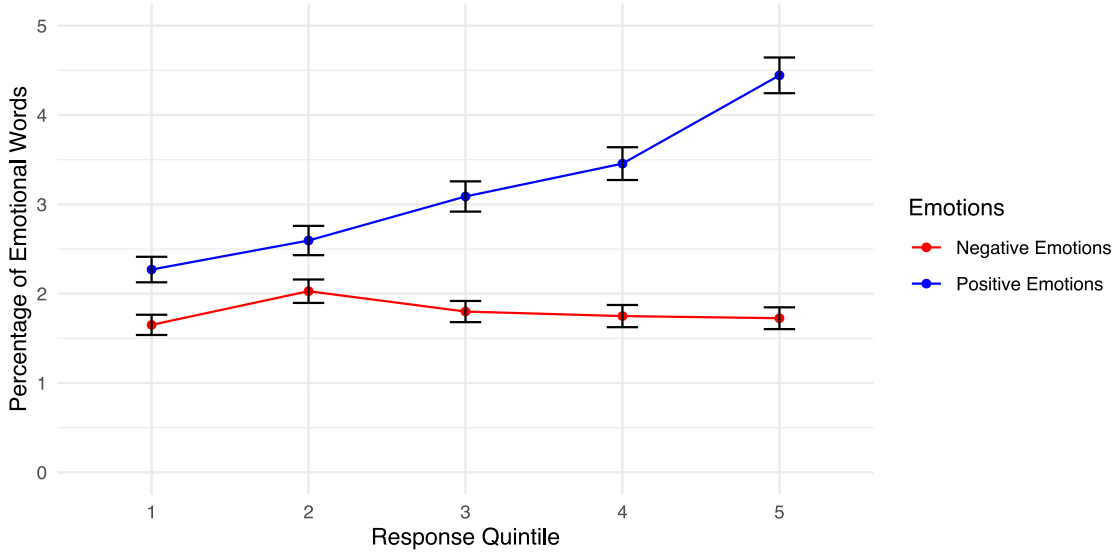
Note. Below diagonal: Within-Individual, Above diagonal: Between-Individual; Pairwise deletion; * p < .05; ** p < .01; *** p < .001.

Emotional Trajectories in Narrative Responses

To investigate the emotional trajectories in narrative responses (RQ1), we computed repeated measures correlations between quintiles and emotions to test whether their trajectories differ from zero slopes. On average, there was a significant correlation between the quintile and positive emotions (r_{rm} = .22, p < .001). But there was no significant correlation between the quintile and negative emotions (r_{rm} = -.01, p = .700), indicating that, on average, the negative emotional trajectory remains relatively flat throughout the narrative responses (see Figure 1).

Figure 1

Positive and Negative Emotional Trajectories Throughout Narrative Responses



Note. N = 375 narrative responses nested in 128 participants. Error bars indicate ± 1 SE.

Effect of Emotional Trajectories on Rater's Evaluations

To investigate the effect of positive and negative emotional trajectories within narrative responses on the rater's evaluations and emotional engagement with the response, we computed six linear mixed-effect regressions (one for each dependent variable; see Table 3).

For competence, the fixed effects accounted for 4.2% of the variance, while the combined fixed and random effects explained 23%. Results showed a small but significant fixed effect related to the sum of negative emotions (B = -0.02, SE = 0.01, t = -2.36, p = .038). This suggests that higher levels of overall negative emotions in responses were associated with decreased competence. In addition, there was a main effect of the slopes of negative emotions (B = -0.18, SE = 0.09, t = -2.12, p = .035), highlighting that increasing negative emotions over the course of the response was associated with lower competence. However, there were no main effects of the sum of positive emotions and the slope for positive emotions. There was no interaction effect of the slopes of positive and negative emotions. Regarding the random effects, the variance for individuals was 0.20 (SD = 0.45), and for questions 0.05 (SD = 0.21), the residual variance was 1.01 (SD = 1.01).

For perceived performance, the fixed effects accounted for 1.3% of the variance, while the combined fixed and random effects explained 62.6%. Results showed a main effect of the slope of negative emotions (B = -0.12, SE = 0.06, t = -2.24, p = .025): Increasing negative emotions throughout the narrative response lowered evaluations of perceived performance. However, there were no main effects of the sum of positive emotions, the sum of negative emotions, and the slope for positive emotion. There was no interaction effect of the slopes of positive and negative emotions. Regarding the random effects, the variance for participants was 0.27 (SD = 0.52), and for questions 0.28 (SD = 0.53), suggesting considerable variation explained by the individuals and the questions. The residual variance was 0.33 (SD = 0.58).

For perceived engagement, the fixed effects accounted for 0.6% of the variance, while the combined fixed and random effects explained 86.4%. Results showed a small but significant main effect of the sum of positive emotions (B = -0.01, SE = 0.00, t = -2.59, p = .010): The more positive emotions present in the overall narrative response, the less engagement was perceived. However, there were no main effect for the sum of negative emotions, both slopes

for positive and negative emotions, and no interaction effect between the slopes for positive and negative emotions. Regarding the random effects, the variance for participants was 0.54 (*SD* = 0.73), and for questions 0.02 (*SD* = 0.14), suggesting considerable variation explained by the individuals. The residual variance was 0.08 (SD = 0.30).

For persuasiveness, the fixed effects accounted for 0.9% of the variance, while the combined fixed and random effects explained 82.2%. Results showed a small but significant main effect of the slope for positive emotions (*B* = 0.05, *SE* = 0.02, *t* = 2.26, *p* = .025): Increasing positive emotions throughout the narrative response increased evaluations of persuasiveness. However, there were no main effects for the sum of positive emotions, the sum of negative emotions, or the slope for negative emotions. There was no interaction effect between the slopes for positive and negative emotions. Regarding the random effects, the variance for participants was 0.54 (*SD* = 0.73), and for questions 0.00 (*SD* = 0.06), suggesting considerable variation explained by the individuals. The residual variance was 0.12 (SD = 0.34).

For rater's emotional engagement with the response, the fixed effects accounted for 1% of the variance, while the combined fixed and random effects explained 56.8%. Results showed no main effects for the sum of positive emotions, the sum of negative emotions, or both slopes for positive and negative emotions. There was no interaction effect between the slopes for positive and negative emotions. Regarding the random effects, the variance for participants was 0.18 (*SD* = 0.42), and for questions 0.19 (*SD* = 0.43), suggesting considerable variation explained by the individuals and the questions. The residual variance was 0.29 (SD = 0.54).

Table 3

Effect of Positive and Negative Emotional Trajectories on Evaluations.

	Competence		**Performance**		**Engagement**		**Persuasiveness**		**Rater's Emotional Engagement**	
Predictors	*B*	SE	*B*	SE	*B*	SE	*B*	SE	*B*	SE
(Intercept)	2.79***	0.14	3.21***	0.31	3.48***	0.11	2.96***	0.08	1.57***	0.26
Sum Pos. Emotions	0.01	0.01	-0.00	0.00	**-0.01****	0.00	-0.00	0.00	-0.00	0.00
Sum Neg. Emotions	**-0.02***	0.01	-0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01
Slope Pos. Emotions	0.08	0.06	-0.04	0.04	0.01	0.02	**0.05***	0.02	0.03	0.03
Slope Neg. Emotions	**-0.18***	0.09	**-0.12***	0.06	-0.05	0.03	-0.04	0.04	-0.04	0.05
Slope Pos. Emotions x Slope Neg Emotions	-0.05	0.08	-0.07	0.05	0.01	0.03	-0.06	0.03	-0.03	0.05
Random Effects										
σ^2	1.01		0.33		0.09		0.12		0.29	
$\tau_{00\ ID}$	0.20		0.27		0.54		0.54		0.18	
$\tau_{00\ Q}$	0.05		0.28		0.02		0.00		0.19	
ICC	0.20		0.62		0.86		0.82		0.56	
Observations	372		351		351		351		351	
Marginal R^2	0.042		0.013		0.006		0.009		0.010	
Conditional R^2	0.230		0.626		0.864		0.822		0.568	

Note. Pos = positive, Neg = negative; Significance level: * p < .05; ** p < .01; *** p < .001

Discussion

Because narratives typically follow emotional trajectories that shape the audience reactions (Del Vecchio et al., 2021; Reagan et al., 2016; Alam & So, 2020; Carrera et al., 2008; Sy et al., 2018), applicants may also emotionally frame their narrative responses to past-behavior questions, ultimately influencing recruiters' evaluations. This study examined emotional framing in narrative responses to past behavior questions to understand how emotional content unfolds in narratives and how it impacts recruiters' evaluations.

Regarding the trajectories of positive and negative emotions in narrative responses (RQ1), on average, positive emotions increased significantly throughout the response, while negative emotions did not. As past-behavior questions naturally invite applicants to describe an initially negative situation they are expected to resolve, an uptick in positive emotional elements would demonstrate a good resolution. Our findings show that participants, *in extenso* applicants, are well aware of the emotional trajectory that showcases their competence to recruiters. When it comes to negative emotions, their flat trajectory may suggest a reluctance to verbally express negative emotional elements during the interview, perhaps to avoid appearing incompetent. Furthermore, negative emotions may be expressed preferably through channels other than words. They might instead manifest themselves in para-verbal (e.g., sighing) or non-verbal (e.g., frowning, smirking) behaviors (App et al., 2011; Simon-Thomas et al., 2009), which were not considered in our operationalization of emotions.

Regarding the effect of emotions and their emotional trajectories in narrative responses on recruiters' evaluations (RQ2), our results revealed that competence – assessed with BARS, decreases when negative emotional content in the overall response is high, and the negative emotions increase throughout the response. Similarly, the performance – assessed on a 5-point scale (1 = *not at all*, 5 = *completely*), decreases when negative emotions increase throughout the response. These findings show that negative emotional content in narrative responses can

impact perceptions of competence and performance. Furthermore, they suggest that this effect applies to different evaluation approaches, both an objective approach (i.e., BARS) and a more subjective and general approach (i.e., a 5-point scale capturing overall impressions). Nevertheless, the impact of negative emotions on competence measured with BARS, typically used as a measure robust to bias (Smith & Kendall, 1963), can be interpreted in two ways. One possibility is that BARS might not fully resist subtle emotional response framing, as responses with increasing negative elements toward the end of the narrative are rated lower, potentially because of the emotional framing itself rather than the actual competence. Alternatively, it may indicate that competent applicants naturally resolve the situations they describe, leading to fewer negative elements in their narratives by the end. To clarify this, future research should compare emotionally framed responses against those that are emotionally neutral.

Regarding other evaluations, engagement was slightly lower when the total positive emotional content in the response was high. However, when positive emotions increased throughout the response, participants were seen as giving more convincing answers. This dual observation evidence two aspects of positive emotions: on the one hand, higher overall positive emotional content in the narrative may suggest that the situation is perceived as globally positive, not requiring the applicant's active involvement in the situation. This could lead recruiters to perceive less engagement. On the other hand, a progressive increase in positive emotions throughout the response may signal that the applicant is overcoming difficulties and thus progressing towards a successful resolution of the situation.

Despite research suggesting that emotional trajectories in narratives would reinforce narrative transportation, the raters' emotional engagement with the narrative response was not affected by the positive or negative emotions within the overall responses nor their trajectories. One possible explanation lies in the format of the narrative. Previous research that found greater transportation in the audience in the presence of emotional shifts in narratives used

written narratives only, where emotional framing is saturated solely in words (Alam & So, 2020; Carrera et al., 2008). In contrast, this study's rater had access to video-recorded narrative responses, integrating verbal, paraverbal, and non-verbal behaviors. Therefore, the rater assessed their emotional engagement using a multimodal narrative response. The multimodality of oral narrative response in our study may have attenuated the effect of emotional trajectories on raters' emotional engagement since positive and negative emotions and their trajectories were operationalized solely based on words.

Implications

Our findings contribute to both research and practice. This research provides evidence that *how a response is told* affects recruiters' perceptions and evaluations. Specifically, recruiters' evaluations, even those grounded in more objective criteria, may be influenced by how applicants framed their responses. First, research should investigate how applicants' communication skills affect interview outcomes, aligning with Huffcutt & Murphy's (2023) call for more research. If communication is not a key competency targeted for the job position, interview outcomes may fail to accurately reflect applicants' actual characteristics, thereby contributing to the recently observed high variability in interview validity (Huffcutt & Murphy, 2023; Sacket et al., 2022). Second, hiring organizations should make their recruiter aware of such potential bias. Furthermore, they should adopt a structured approach for assessing responses to guarantee greater objectivity. For example, organizations could consider relying on more than one recruiter's evaluation to mitigate irrelevant inferences and biases in evaluations (see Campion et al., 1997).

Limitations and Future Research

Our study has several limitations that should be addressed in future research to investigate emotional framing in job interviews further. One limitation concerns the sample used in this study. We used narrative responses collected as part of an experiment in which participants

answered past-behavior questions in a lab setting. Their responses may not reflect those of applicants involved in real, high-stakes interviews, although participants were incentivized to perform through progressive performance-based remuneration. This could limit the generalizability of our results. Therefore, future studies should replicate our findings with a more diverse sample and across job positions.

A second limitation of our study is that we used the emotional valences of words within narrative responses. However, emotions are multimodal; they manifest in non-verbal (e.g., facial expressions, gestures) and para-verbal (e.g., intonations, pitch) behaviors that can vary to indicate emotions and intensities (App et al., 2011; Simon-Thomas et al., 2009). Future research should investigate the role of emotional framing in responses by adopting a multimodal analysis of emotion and considering non-linear relationships. One promising avenue would be to use artificial intelligence to detect verbal, non-verbal, and para-verbal emotional behaviors in applicants' responses that affect recruiters' evaluations. For instance, a system may be developed to process video recordings of applicants' responses while having access to recruiters' evaluations. Using different algorithms, such a system could identify patterns in verbal (e.g., words), non-verbal (e.g., facial expressions), and paraverbal (e.g., pitch) behaviors that contribute to higher evaluations. This approach would help identify emotional behaviors and quantify their effect on evaluations.

Conclusion

Emotional trajectories within narratives shape audience reactions, yet this raises the question of how emotional framing of narrative responses in job interviews influences recruiters' evaluations of applicants' competence, performance, persuasiveness, engagement, and their emotional engagement with the response. Our findings reveal that positive and negative emotional trajectories impact some of these evaluations. Because very little research exists on applicants' communication skills and their effect on recruiters' evaluation, future research

should continue to explore this by adopting a multimodal approach and investigating how this relates to interview validity.

References

Alam, N., & So, J. (2020). Contributions of emotional flow in narrative persuasion: An empirical test of the emotional flow framework. *Communication Quarterly*, *68*(2), 161-182. https://doi.org/10.1080/01463373.2020.1725079

App, B., McIntosh, D. N., Reed, C. L., & Hertenstein, M. J. (2011). Nonverbal channel use in communication of emotion: how may depend on why. *Emotion*, *11*(3), 603-617. https://doi.org/10.1037/a0023164

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 456. https://doi.org/10.3389/fpsyg.2017.00456

Bakdash, J. Z., & Marusich, L. R. (2024). *rmcorr: Repeated Measures Correlation*. In R package version 0.7.0. https://CRAN.R-project.org/package=rmcorr

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. *Journal of Business and Psychology*, *29*(4), 593-604. https://doi.org/10.1007/s10869-014-9350-0

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*. https://www.liwc.app

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, *6*(32), eaba2196. https://doi.org/10.1126/sciadv.aba2196

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. *European Journal of Work and Organizational Psychology*, *29*(3), 330-343. https://doi.org/10.1080/1359432X.2019.1704265

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655-702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Carrera, P., Caballero, A., & Muñoz, D. (2008). Comparing the effects of negative and mixed emotional messages on predicted occasional excessive drinking. *Substance Abuse: Research and Treatment*, *1*, 1-7. https://doi.org/10.1177/117822180800100001

Del Vecchio, M., Kharlamov, A., Parry, G., & Pogrebna, G. (2021). Improving productivity in Hollywood with data science: Using emotional arcs of movies to drive product and service innovation in entertainment industries. *Journal of the Operational Research Society*, *72*(5), 1110-1137. https://doi.org/10.1080/01605682.2019.1705194

Germanier, E., Bangerter, A., Orji, K., Schmid Mast, M., Renier, L. A., He, M., & Garner, P. P. (2024). *Responses to Past-Behavior Questions in Face-to-Face and Asynchronous Video Interviews: Storytelling, Interview Performance and Criterion Validity [Manuscript submitted]*. Institute of Work and Organizational Psychology, University of Neuchâtel.

Green, M. C., & Appel, M. (2024). Narrative transportation: How stories shape how we see ourselves and the world. *Advances in Experimental Social Psychology*, *70*(1), 1-82. https://doi.org/10.1016/bs.aesp.2024.03.002

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701-721. https://doi.org/10.1037/0022-3514.79.5.701

Huffcutt, A. I., Howes, S. S., Murphy, D. D., & Murphy, S. A. (2024). Enhancing consistency of maximal responding in behavior description interviews: An exploration of priming and response length. *Personnel Assessment and Decisions*, *10*(1), 1-11. https://doi.org/10.25035/pad.2024.01.001

Huffcutt, A. I., & Murphy, S. A. (2023). Structured interviews: moving beyond mean validity…. *Industrial and Organizational Psychology*, *16*(3), 344-348. https://doi.org/10.1017/iop.2023.42

Imada, A. S., & Hakel, M. D. (1977). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology*, *62*(3), 295-300. https://doi.org/10.1037/0021-9010.62.3.295

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158-168). Sage Publications.

Kessler, R. (2006). *Competency-based interviews: Master the tough new interview style and give them the answers that will win you the job*. Frankilin Lakes : Career Press.

Krause, R. J., & Rucker, D. D. (2020). Strategic storytelling: When narratives help versus hurt the persuasive power of facts. *Personality and Social Psychology Bulletin*, *46*(2), 216-227. https://doi.org/10.1177/0146167219853

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), 1-15. https://doi.org/10.1016/j.hrmr.2020.100789

Mandelbaum, J. (2013). Storytelling in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 492-507). Chichester, UK: John Wiley & Sons.

Nabi, R. L. (2015). Emotional flow in persuasive health messages. *Health Communication*, *30*(2), 114-124. https://doi.org/10.1080/10410236.2014.974129

Orji, K., Bangerter., A., Germanier, E., Renier, L. A., Schmid Mast, M., H., M. , & Garner, P. N. (2024). *Extended Responses in Asynchronous Video Interviews: Investigating Frequency, Content, and Interview Outcomes [Manuscript in preparation]*. Institute of Work and Organizational Psychology, University of Neuchâtel.

Piolat, A., Booth, R., Chung, C. K., Davids, M., & Pennebaker, J. (2011). La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. *56*(3), 145-159. https://doi.org/https://doi.org/10.1016/j.psfr.2011.07.002

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ralston, S. M., Kirkwood, W. G., & Burant, P. A. (2003). Helping interviewees tell their stories. *Business Communication Quarterly*, *66*(3), 8-22. https://doi.org/10.1177/108056990306600303

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, *5*(1), 1-12.

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. In R package version 2.4.6. https://CRAN.R-project.org/package=psych

Roulin, N., Bangerter, A., & Wüthrich, U. (2012). *Réussir l'entretien d'embauche comportemental: La méthode pour identifier et sélectionner les futurs employés performants*. Bruxelles: De Boeck Professionals.

Silvester, J. (1997). Spoken attributions and candidate success in graduate recruitment interviews. *Journal of Occupational and Organizational Psychology*, *70*(1), 61-73. https://doi.org/10.1111/j.2044-8325.1997.tb00631.x

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: evidence from vocal burst displays. *Emotion*, *9*(6), 838-846. https://doi.org/10.1037/a0017810

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*(2), 149-155. https://doi.org/10.1037/h0047060

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, *80*(5), 587-606. https://doi.org/10.1037/0021-9010.80.5.587

Sy, T., Horton, C., & Riggio, R. (2018). Charismatic leadership: Eliciting and channeling follower emotions. *The Leadership Quarterly*, *29*(1), 58-69. https://doi.org/10.1016/j.leaqua.2017.12.008

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology*, *75*(3), 277-294. https://doi.org/10.1348/096317902320369712

Van Laer, T., De Ruyter, K., Visconti, L. M., & Wetzels, M. (2014). The extended transportation-imagery model: A meta-analysis of the antecedents and consequences of

consumers' narrative transportation. *Journal of Consumer Research*, *40*(5), 797-817. https://doi.org/10.1086/673383

Wasielewski, P. L. (1985). The emotional basis of charisma. *Symbolic Interaction*, *8*(2), 207-222.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Winkler, J. R., Mengelkamp, C., & Appel, M. (2022). Real-time responses to stories: Linking valence shifts to post-exposure emotional flow and transportation. *Communication Research Reports*, *39*(5), 237-247. https://doi.org/10.1080/08824096.2022.2119380

II Appendix 2: Study 2

Germanier, E., Bangerter, A., Renier, L. A., Kleinlogel, E. P., Schmid Mast, M., Jayagopi, D. B., Shubham, K., & Roulin, N. (2023). *Effects of Interview Medium and Culture on Applicant Storytelling, Disfluencies and Evaluations in Behavioral Interviews* [Unpublished manuscript]. Institute of Work and Organizational Psychology, University of Neuchâtel.

Status: The manuscript was submitted to a journal but was rejected. The version in this thesis is the manuscript initially submitted, as revisions are currently on hold.

Effects of Interview Medium and Culture on Applicant Storytelling, Disfluencies and Evaluations in Behavioral Interviews

Elisabeth Germanier[1], Adrian Bangerter[1], Laetitia A. Renier[2], Emmanuelle P. Kleinlogel[3], Marianne Schmid Mast[2], Dinesh Babu Jayagopi[4], Kumar Shubham[4], Nicolas Roulin[5]

[1] Institute of Work and Organizational Psychology, University of Neuchâtel, Switzerland

[2] Department of Organizational Behavior, University of Lausanne, Switzerland

[3] CEMOI Laboratory, IAE Reunion, University of Reunion Island, Saint-Denis, France

[4] International Institute of Information Technology, Bangalore, India

[5] Department of Psychology, Saint Mary's University, Canada

This research was supported by SNSF grant 10521C_197479, by the Faculty of Business and Economics of the University of Lausanne and by the Machine Intelligence and Robotics (MINRO) center at IIIT Bangalore, India.

Correspondence concerning this article should be addressed to Elisabeth Germanier, Institute of Work and Organizational Psychology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. E-Mail: elisabeth.germanier@unine.ch.

The authors have no conflicts of interest to declare.

Abstract

Asynchronous Video Interviews (AVIs) introduce new challenges for applicants and recruiters because of the lack of real-time interaction, for instance using past-behavior questions inviting applicants to recount work-related situations. This experiment (n = 299) investigated the impact of interview medium (face-to-face (FTF) interviews vs. AVIs with avatar questions (AQ) vs. AVIs with written questions (WQ)) and culture (Swiss vs. Indian) on applicant responses (storytelling, disfluencies) and raters' evaluations (applicant perceived engagement and perceived self-confidence) in mock behavioral interviews. Participants' speech fluency was comparable across cultures, with minor differences across interview media. AQ and WQ interviews featured more storytelling responses than FTF interviews and Swiss participants produced more storytelling than Indian participants. Raters' evaluations varied by interview medium and culture.

Keywords: Asynchronous video interview; past-behavior question; culture differences; applicant behaviors; storytelling.

Practitioners' Points:

- Applicants tend to prefer face-to-face (FTF) interviews over Asynchronous Video Interviews (AVIs).
- Culture influences applicants' and recruiters' expectations of appropriate behaviors.
- Applicant speech fluency and raters' evaluation of applicants perceived self-confidence is similar across the interview media.
- AVIs with written questions and AVIs with avatar questions are comparable in terms of applicant responses and raters' evaluation.
- Culture influences applicant storytelling and how raters perceive applicants' engagement and self-confidence.

Effects of Interview Medium and Culture on Applicant Storytelling, Disfluencies and Evaluations in Behavioral Interviews

Introduction

Behavioral interviewing is a best practice in personnel selection (Campion et al., 1997). This type of structured interview is characterized by situational and past-behavior questions (Motowidlo, 1999). Past-behavior questions invite applicants to describe their actions in a past work-related situation for recruiters to assess their mastery of competencies (Janz, 1982). Applicants should tell a story explaining how they faced and resolved a problematic situation while portraying themselves advantageously (Bangerter et al., 2014; Stevens & Kristof, 1995).

Increasingly, selection interviews, including behavioral ones, are computer-mediated. An example is Asynchronous Video Interviews (AVIs), where applicants record themselves answering questions on an online platform, without interacting with a recruiter (Basch et al., 2020; Lukacik et al., 2022). The growing use of AVIs, exemplified by HireVue surpassing 33 million worldwide interviews in 2022 (Hirevue, 2022), outpaces current understanding of their effects on the interview process. Research has reported more negative applicant attitudes towards AVIs compared to face-to-face (FTF) interviews. Applicants perceive less chance to perform well in AVIs and report lower performance in AVIs compared to FTF interviews (Basch et al., 2020; Kleinlogel et al., 2023). Nevertheless, applicants' performance in FTF interviews may be similar to AVIs when they have no preparation time and are evaluated via videotaped assessments (Kleinlogel et al., 2023). Yet, our understanding of AVIs remains limited. Considering the lack of recruiter presence and the variety of possible designs (Lukacik et al., 2022), little is known about applicant responses in AVIs (vs. FTF interviews), particularly concerning storytelling and speech fluency. Furthermore, the increasing use of AVIs across the globe highlights a research gap concerning cross-cultural differences in applicant responses and evaluations across various cultures and interview media. Since culture

underlies social norms and rules (Liu et al., 2022), applicant responses may differ across cultures. Also, recruiters may be biased in their evaluation of applicants from different cultures (Arseneault & Roulin, 2023).

In this article, we seek to gain insights into the impact of two AVI designs, specifically those featuring avatar questions (AQ) and written questions (WQ), on applicant responses and recruiters' evaluations in behavioral interviews compared to FTF interviews. Additionally, we aim to better understand how culture affects both applicants and recruiters across interview media. This study employs a between-subjects design to compare FTF interviews with AQ and WQ interviews (using data from Kleinlogel et al., 2023), investigating applicant responses and recruiters' evaluations, along with cross-cultural differences.

This research fills a gap in our understanding of how the interview medium can influence applicant responses and recruiters' evaluations. It also has important practical implications for organizations considering using, or already engaged in AVIs. Comparing AQ and WQ to FTF interviews might help making informed decisions when designing selection interviews. Furthermore, understanding cross-cultural differences can contribute to the development of more inclusive and fairer selection interviews that consider cultural diversity.

Behavioral Interviews

Behavioral interviews are one type of structured interviews that involve questions about past work situations (e.g., "Tell me about a situation where you had to deal with an angry client"; Roulin et al., 2012). Responses to past-behavior questions are valid predictors of job performance (Day & Carroll, 2003; Hartwell et al., 2019). Their predictive validity is based on the assumption that past job performance predicts future job performance (Janz, 1982). By inviting applicants to talk about how they behaved in a past work-related situation, recruiters can assess applicants' competencies (Levashina et al., 2014; Motowidlo, 1999; Roulin et al., 2012).

To answer past-behavior questions, applicants should tell a story. This entails describing a past work-related situation and the actions undertaken to solve it (Ralston et al., 2003). To provide complete stories, applicants should ideally follow the STAR model, describing the initial Situation, their Tasks, the Actions they performed, and the Results obtained (Kessler, 2006). Applicants responding to past-behavior questions with stories are perceived as more hirable. Nonetheless, they often fail to produce stories, mostly resorting to generic descriptions of past situations or statements about their values or personal opinions (Bangerter et al., 2014). When they do produce stories, they focus on describing situations instead of their actions and the results obtained (Bangerter et al., 2014).

Beyond content, storytelling performance also matters. Fluent speech conveys a positive impression of being more confident, competent, assertive, credible, and convincing (Burgoon et al., 1990; Leigh & Summers, 2002). In the question-answer dynamic of selection interviews, applicants should initiate answers rapidly after the recruiter's question (Brosy et al., 2016). However, they might need to delay their answer to prepare it, organize the next part of their speech or to repair an utterance when a problem (e.g., syntactic, lexical, or phonological; Brennan, 2000) is detected. But disfluent speech (Fox Tree, 1995) creates a negative impression of being hesitant and unconfident (Brennan & Williams, 1995; Corley & Stewart, 2008), thus negatively affecting recruiter evaluations (Brosy et al., 2016).

Along with speech fluency, recruiters' overall impressions of applicant behaviors also affect evaluations. As recruiters evaluate applicant responses, they infer characteristics like personality traits and social skills (DeGroot & Gooty, 2009; Frauendorfer & Schmid Mast, 2015; Huffcutt et al., 2001; Mino, 1996). Recruiters evaluate applicants more positively when they appear more confident and motivated (Gifford et al., 1985; Tessler & Sushelsky, 1978). In contrast, recruiters rate applicants lower if they exhibit signs of anxiety, perceiving them as less assertive and less interpersonally warm (Feiler & Powell, 2016; Finnerty et al., 2016; Vijay

et al., 2021). Yet, most studies have focused on applicant responses and recruiter evaluations in FTF interview settings (Bangerter et al., 2014; Brosy et al., 2016; DeGroot & Motowidlo, 1999; Forbes & Jackson, 1980; Imada & Hakel, 1977; Nguyen & Gatica-Perez, 2015; Parsons & Liden, 1984) and we currently lack understanding about the content of applicant responses and how they impact recruiter evaluations in AVIs.

Asynchronous Video Interviews

Technology-mediated interviews can be divided into two categories: synchronous (e.g., telephone, videoconference) and asynchronous (e.g., AVIs). The key distinction between these categories is whether a recruiter is present during the interview or not. In synchronous interviews, applicants remotely engage in real-time interaction with recruiters whereas they are deprived of it in AVIs (Lukacik et al., 2022). Instead, they record their responses to pre-determined questions on a platform, either asked by an avatar (avatar questions, AQ) or via written questions (WQ) displayed on-screen, and their answers are evaluated later (Lukacik et al., 2022). While AVIs offer flexibility in time and location (Basch & Melchers, 2019), applicants perceive lower chance of performing in comparison to other interview media, which can lead to lower self-rated performance (Basch et al., 2020; Kleinlogel et al., 2023; Langer et al., 2017). But Kleinlogel et al. (2023) found that despite applicant lower self-rated performance, their other-rated performance in AVIs was similar to FTF interviews when consistent rating procedures were applied across both interview media and when preparation time was removed in AVIs.

Applicant perceptions of poor performance in AVI settings may be explained by the four media attributes from Potosky's (2008) framework (see Kleinlogel et al., 2023). First, interactivity (i.e., the rate of turn exchange between interlocutors) is inexistent due to the lack of recruiters' presence. Second, social bandwidth (i.e., the number of verbal and nonverbal cues conveyed by the medium) is limited not only for applicants because of computer-based

constraints, but also for recruiters who cannot transmit any cues themselves (Daft & Lengel, 1986; Potosky, 2008; Van Iddekinge et al., 2006). Third, the transparency of the medium (i.e., the extent to which interlocutors are not aware of the medium while interacting) is reduced because applicants are reminded that the computer is an obstacle between themselves and recruiters. Fourth, the sense of surveillance (i.e., whether a third party can enter or monitor the conversation), is higher because there is little assurance of videorecording confidentiality.

Due to the lack of two-way interaction in AVIs, differences in applicant storytelling responses to past-behavior questions are likely to emerge compared to FTF interviews. It is plausible that applicant may tell more detailed stories in AVIs, as they might find it easier to provide extended answers, unless their response time is severely restricted. This differs from FTF interviews, where effective organization of speech turns and management of speaking time can be more challenging (Brosy et al., 2016). Alternatively, applicants might share fewer stories since storytelling is an interactive process that involves at least one engaged listener (Mandelbaum, 2013), and distracted listeners are detrimental to storytelling quality (Bavelas et al., 2000). Given these potential differences in storytelling behaviors between AVIs and FTF interviews, an examination of how various AVI designs affect storytelling responses as compared to traditional FTF interviews is warranted:

Research Question 1. What effect does the interview medium have on interviewees' storytelling responses to past-behavior questions?

The absence of a recruiter also raises questions about applicants' speech fluency. In AVIs, applicants are deprived of recruiters' real-time feedback and nonverbal cues typically available in FTF interviews. They thus have to monitor their responses without the immediate benefit of external validation or correction. In addition, they must strive to maintain a natural flow of speech. This new challenge necessitates an investigation into how different AVI

designs, as compared to FTF interviews, affect interviewees' disfluencies in response to both past-behavior questions and other question types:

Research Question 2. What effect does the interview medium have on interviewees' disfluencies in behavioral interviews?

The constraints of a medium can also influence how recruiters assess applicant engagement and self-confidence. Limited social bandwidth and absence of direct interaction hinder applicants from using gestures or immediacy behaviors such as eye contact, smiling and managing interpersonal distance (Imada & Hakel, 1977). These behaviors positively impact recruiters' evaluation of interviewee self-confidence and motivation in FTF interviews (Gifford et al., 1985; Tessler & Sushelsky, 1978). Since recruiters cannot observe these cues in AVIs, their evaluations may differ from those in FTF interviews. While initial research indicates that applicants are perceived as less stressed in AQ and WQ than in FTF interviews (Kleinlogel et al., 2023), there remains a gap in our understanding of how recruiters assess applicant engagement and self-confidence across different AVI designs in comparison to FTF interviews:

Research Question 3. What effect does the interview medium have on raters' evaluations of interviewees?

Cultural Differences in Selection Interviews

Culture is a set of attitudes, values and beliefs that influence one's behaviors and expectations (Manroop et al., 2013). According to Hofstede (1984, 2010), a country's culture can be understood through six dimensions: 1) power distance (i.e., how members of a culture deal with social inequality), 2) individualism (i.e., how members of a culture relate to the community), 3) uncertainty avoidance (i.e., how members of a culture avoid uncertainty), 4) masculinity (i.e., whether the goals shared by members of a culture are masculine with work as a central aspect), 5) long-term orientation (i.e., how members of a culture are committed to past traditions) and 6) indulgence (i.e., how members of a culture tolerate human natural drives;

Triandis, 1982). Cultural dimensions underlying social norms and rules shape the verbal and nonverbal behaviors of its members (Liu et al., 2022).

In selection interviews, culture influences the representation of appropriate behaviors and strategies for positive image creation. Depending on the culture, applicants showing signs of excitement during interviews are more or less appreciated by recruiters than calmer applicants (Bencharit et al., 2019). Also, applicant strategies to build a positive image of themselves vary across cultures (Arseneault & Roulin, 2021; Sandal et al., 2014). For example, applicants from performance-oriented cultures are more likely to promote themselves than applicants from cultures that value performance less (Sandal et al., 2014). While storytelling is considered a good way to create a positive impression (Stevens & Kristof, 1995), it is unclear how storytelling is used in response to past-behavior questions across cultures. In addition, culture also influences individual speech fluency. Some cultures are more prone to consider fast talk with short speaking turns and pauses as positive in general (Stenström, 2011), but cultural differences in tolerance of speech disfluencies in selection interviews are unexplored. In this light, there is a need to investigate applicant responses across cultures.

Considering cultural differences across interview media, questions also arise about how the interview medium impacts applicant responses and recruiters' evaluations across cultures. For example, AVIs' limited social bandwidth may challenge applicants from cultures accustomed to expansive hand gestures (Kita, 2009) because they cannot employ these gestures in AVIs as they would in FTF interviews. Moreover, cultural attitudes towards technology may also play a role. Cultures that are more open to technology adoption (Jan et al., 2022) may have applicants and recruiters who are more familiar with computer-mediated interaction. This familiarity may result in different responses and evaluations in AVIs when compared to applicants from cultures that are less technology friendly. Furthermore, cultural differences between recruiters and applicants may affect recruiters' evaluations. Notably, greater cultural

similarity leads to better recruiters' evaluations of interview performance (Arseneault & Roulin, 2023). However, it is worth noting that Kleinlogel et al. (2023), with Swiss raters, found no difference in perceived stress and performance among applicants from two different cultures (i.e., Indian and Swiss applicants). Despite initial research (Arseneault & Roulin, 2023; Kleinlogel et al., 2023), there is still a lack of understanding of differences in applicant responses and evaluations across various cultures and interview media:

Research Question 4. What effect does culture have on interviewees' responses and raters' evaluations of interviewees across interview media?

To answer our research questions, we used data from an experiment originally conducted by Kleinlogel et al. (2023). In the current study, their data were transcribed and coded for interviewees' storytelling, disfluencies, as well as rated for interviewee perceived engagement and self-confidence.

Method

Participants

The sample consisted of 299 male students (M_{age} = 22.18, *SD* = 3.07), split between Swiss nationality (49.5%) and Indian nationality (50.5%). In terms of their educational background, 40.8% were in Bachelor's programs, 54.9% were in Master's programs and 2% pursued other studies (e.g., PhD). Compensation varied between data location: Swiss participants (*N* = 148) were paid CHF 15 (about USD 16.50), while Indian participants (*N* = 151) were paid RS 50 (about USD 0.60). These incentives matched the minimum wage in participants' countries.

Procedure

Overview

The study consisted in one session lasting about 30 minutes held in a lab. Participants were randomly assigned to one of the three conditions (interview medium: FTF vs. AQ vs.

WQ). First, they read and signed a consent form. Then, they participated in a mock job interview according to the condition they were assigned to. The job interview stage comprised three steps: (a) participants indicated the job position for which they would like to apply for, (b) they were instructed to behave, during the job interview, as if they were applying for the selected job position, and (c) they performed the job interview for around 10 minutes. Participants were video recorded during the job interview using one camera to obtain a frontal view recording.

Job Interview Content

The job interview comprised five questions, among which the first three questions were the same for all participants. The first question asked participants to introduce themselves and describe their current situation. The second and third questions were past behavior questions (i.e., "describe a situation in which you had to manage several tasks or projects at the same time and how you handled this situation" and "give an example of a situation in which you took the initiative to get things done and in which you were successful"). The fourth and fifth questions, randomly selected from a pool of 20 questions, were more general typical job interview questions that varied across participants (e.g., questions about personality, why they should be hired).

Job Interview Conditions

In the FTF condition, the experimenter asked participants to sit at a table and wait for the "recruiter" (another experimenter) to join them, and then left the room. The recruiter started the interview by asking the first question. Each time the applicant replied, the recruiter answered, "Thank you for your answer." and moved to the next question. If the participant did not answer at all, the recruiter said "Ok, I will ask you the next question." Recruiters were trained to follow the same script of interview questions to ensure high standardization. Also, to look as natural as possible, recruiters were allowed some backchanneling (mainly head

nodding), to take notes during the interview, and to reply concisely if participants asked questions.

In both AQ and the WQ conditions, the experimenter opened a software interface developed for the study. Participants in the AQ condition interacted with an avatar that had the same appearance and voice as the country-specific FTF recruiters who asked the interview questions. Participants in the WQ condition saw the interview questions in a written format only. In both conditions, participants were required to answer each question without any prior preparation time or time limitations for their responses, and they could proceed to the next question by clicking the 'next' button.

Measures

Storytelling

For each participants' answer to a past-behavior question, four aspects of storytelling were coded before being averaged across questions. If participants told more than one story in their answer, coding was only performed on the first story mentioned. Two Swiss independent coders were involved in these measures so that one coder was responsible for the Swiss data and the other for the Indian data. To assess reliability, we initially selected 40 videos for the Swiss data and 40 for the Indian data. Due to some videos being unreadable, reliability was established on double-coding of 34 Swiss videos and 39 Indian videos with a second coder.

Stories. This was coded as "1" when participant's answer featured a story and coded as "0" otherwise. The interrater reliability for the stories was $r_{CH} = .60$ and $r_{IND} = .93$ ($p < .01$).

STAR Situation. We coded references to five situational elements: "when" (temporal markers), "where" (spatial markers), "actor 1" and "actor 2" (who was involved, i.e., collaborators or the employer), and finally "problem" (what hinders or causes the action). Each situational element was coded as "1" if they occurred once or more and coded as "0" otherwise. The STAR Situation score was then computed as the sum of all situational elements, ranging

thus from 0 (= *no element mentioned*) to 5 (= *all elements were mentioned*). The interrater reliability for the STAR Situation was r_{CH} = .61 and r_{IND} = .98 (p < .01).

STAR Task-Action. We coded references to four task-action related elements: "what" (the task, what was expected), "what for" (the action, what was done/planned linked to the objectives, the work done, and the efforts provided), "planification/organization" (presence of conjugated verbs referring to planning, coordination, or communication), and "how" (presence of conjugated action verbs). Each task-action element was coded as "1" if they occurred once or more and coded as "0" otherwise. The STAR Task-Action score was then computed as the total sum of all task-action elements, ranging thus from 0 (= *no element mentioned*) to 4 (= *all elements were mentioned*). The interrater reliability for the STAR Task-Action was r_{CH} = .55 and r_{IND} = .95 (p < .01).

STAR Result. We coded references to four result-related elements: "general success" (indication of success/failure), "specific success" (indication of means to measure success/failure), "effectiveness" (indication of agreement with/between actors, evolution of the problem, or potential improvement), "impression of others" (reactions of others to the story). Each result element was coded as "1" if they occurred once or more and coded as "0" otherwise. The STAR Result score was then computed as the total sum of result elements, ranging thus 0 (= *no element mentioned*) to 4 (= *all result elements were mentioned*). The interrater reliability for the STAR Result was: r_{CH} = .64 and r_{IND} = .94 (p < .01).

Disfluencies

For each participants' answer to the five selection interview questions, we measured two kinds of disfluencies: fillers (i.e., *uh* or *um*) and repetitions (i.e., repetitions of words or part of sentences, when participants interrupted an utterance and then partially or completely repeated them). These measures were first assessed for each interview question separately and then summed to calculate the total number of fillers and repetitions per interview. We then

computed rates of occurrence of filler words and repetitions per hundred words produced by the participants when answering questions. One person coded both measures for the Swiss and Indian participants. Reliability was established by double-coding 20 transcribed interviews for the Swiss data and 20 for the Indian data with a second coder. The interrater reliability for fillers was r_{CH} =.99 and r_{IND} = .99 ($p < .01$) and for repetitions r_{CH} = .99 and r_{IND} = .99 ($p < .01$).

Raters' Evaluations of Participants

For each participants' answer to the five selection interview questions (i.e., both behavioral questions and others), perceived engagement and perceived self-confidence were coded before being averaged across questions. Two Swiss independent raters were involved in these measures so that one rater was responsible for the Swiss data and the other for the Indian data. Reliability was established by double-coding 30 videos for the Swiss data and 30 for the Indian data with a second rater.

Perceived Engagement. We coded for perceived engagement using one item. The two Swiss raters evaluated from 1 (= *not at all*) to 5 (= *totally*) to what extent the participant appeared engaged while responding to the question. The raters evaluated perceived engagement from 1 (= *not at all*) to 5 (= *totally*). The interrater reliability for perceived engagement was r_{CH} = .92, and r_{IND} = .78 ($p < .01$).

Perceived Self-Confidence. We coded for perceived self-confidence using one item. The two Swiss raters evaluated from 1 (= *not at all*) to 5 (= *totally*) to what extent the participant appeared confident during the selection interview. The interrater reliability for perceived self-confidence was r_{CH} = .75 and r_{IND} = .78 ($p < .01$).

Results

Due to the unavailability of some data (e.g., unreadable videos), a sample of N = 289 was available for analyses. Descriptive statistics for dependent variables by condition and by

culture are presented in Table 1. Descriptive statistics and correlations for the main variables are presented in Table 2.

Table 1.

Descriptive Statistics of Dependent Variables per Interview Medium and Culture.

	Interview Medium									Culture					
	FTF			AQ			WQ			Swiss			Indian		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Stories	96	0.77	0.32	98	0.88	0.26	95	0.86	0.30	141	0.88	0.27	148	0.80	0.31
STAR Situation	96	2.05	1.12	98	2.21	0.97	95	2.30	1.17	141	2.58	0.91	148	1.81	1.11
STAR Task-Action	96	1.97	0.99	98	1.89	0.91	95	2.09	0.98	141	2.44	0.75	148	1.55	0.94
STAR Result	96	0.74	0.90	98	0.77	0.66	95	1.01	0.84	141	0.95	0.84	148	0.73	0.77
Fillers	96	2.60	1.83	98	2.69	1.82	95	2.74	1.97	141	2.62	1.78	148	2.72	1.95
Repetitions	96	2.36	1.32	98	2.90	1.34	95	2.44	1.35	141	2.58	1.47	148	2.57	1.24
Perceived Engagement	96	3.42	1.08	98	3.10	1.12	95	3.24	1.12	141	2.43	0.84	148	4.03	0.69
Perceived Self-Confidence	96	3.36	0.85	98	3.44	0.67	95	3.53	0.69	141	3.09	0.55	148	3.78	0.74

Notes. FTF = face-to-face interviews; AQ = AVIs with avatar questions; WQ = AVIs with written questions; STAR = Situation-Task-Action-Result.

Table 2.

Descriptive Statistics and Correlations for the Main Variables.

		n	*M*	*SD*	1	2	3	4	5	6	7
1	Stories	289	0.84	0.29							
2	STAR Situation	289	2.19	1.09	0.62 **						
3	STAR Task-Action	289	1.98	0.96	0.52 **	0.64 **					
4	STAR Result	289	0.84	0.81	0.38 **	0.42 **	0.47 **				
5	Fillers	289	2.67	1.87	-0.04	-0.05	-0.13 *	-0.10			
6	Repetitions	289	2.57	1.35	-0.07	-0.07	-0.11	-0.06	0.23 **		
7	Perceived Engagement	289	3.25	1.11	0.06	-0.03	-0.07	0.14 *	-0.06	-0.09	
8	Perceived Self-Confidence	289	3.44	0.74	0.15 **	0.03	-0.01	0.16 **	-0.09	-0.11	0.63 **

Notes. Correlations computed with pairwise deletion; STAR = Situation-Task-Action-Result; significance level * $p < .05$, ** $p < .01$.

Effect of Interview Medium and Culture on Storytelling (STAR)

To investigate the effect of interview medium and culture on participants production of stories, we conducted a three-way loglinear analysis to assess the relationships between these variables, including their main and interaction effects. We used backwards elimination with a significance of $p < .05$ to identify which terms would be included in the final model. The analysis indicated no highest-order interaction (i.e., interview medium x culture x stories; $\chi^2(4) = 4.65$, $p = .36$) but significant two-way interactions (i.e., interview medium x stories, culture x stories; $\chi^2(12) = 28.24$, $p < .01$). Overall, the final model including the two interactions fitted the data well, $\chi^2(6) = 5.78$, $p = .45$.

To understand the respective effects of the interview medium and culture on participants' productions of stories in more detail, we computed cross tabulations and two chi-square tests (Field, 2013). See Table 3. The first chi-square test revealed a significant relationship between interview medium and the frequency of interviewees' production of stories, $\chi^2(4) = 15.42$, $p < .01$. Regarding the effect of the interview medium, 7.3% of interviewees in the FTF condition produced no stories in response to both past-behavior questions, 32.3% produced one storytelling response, whereas 60.4% produced stories in response to both past-behavior questions. In the AQ condition, 4.1% of interviewees produced no stories in response to both past-behavior questions, 15.3% produced one storytelling response, whereas 80.6% produced stories in response to both past-behavior questions. In the WQ condition, 7.4% of interviewees produced no stories in response to both past-behavior questions, 12.6% produced one storytelling response, whereas 80% produced stories in response to both past-behavior questions. This suggests that storytelling responses to both past-behavior questions were more frequent in both AQ and WQ interviews than in FTF interviews.

The second chi-square test revealed a significant relationship between culture and the frequency of participants' production of stories, $\chi^2(2) = 7.37$, $p = .03$. Among Swiss interviewees, 5% produced no stories to both past-behavior questions, whereas 14.2% produced one storytelling response, and 80.8% produced stories for both past-behavior questions. Among Indian interviewees, 7.4% produced no stories to both past-behavior questions, whereas 25.7% produced one storytelling response, and 66.9% produced stories for both past-behavior questions. These observations indicate that storytelling responses to both past-behavior questions were more frequent among Swiss interviewees than their Indian counterparts.

Table 3.

Contingency Table for Stories by Interview Medium and Culture.

		Stories in the two past-behavior questions						
		No storytelling responses		**1 storytelling response**		**2 storytelling responses**		
Interview Medium	**Culture**	*n*	%	*n*	%	*n*	%	Total count
FTF	Swiss	4	8.3%	13	27.1%	31	64.6%	48
	Indian	3	6.3%	18	37.5%	27	56.2%	48
AQ	Swiss	2	4.1%	3	6.1%	44	89.8%	49
	Indian	2	4.1%	12	24.5%	35	71.4%	49
WQ	Swiss	1	2.3%	4	9.1%	39	88.6%	44
	Indian	6	11.8%	8	15.7%	37	72.5%	51

Notes. FTF = face-to-face interviews; AQ = AVIs with avatar questions; WQ = AVIs with written questions.

To investigate participants' production of STAR narrative elements, we first checked whether there were effects of the interview medium and culture on interviewees' response length. The two-way ANOVA indicated a main effect of culture, $F(1,283) = 9.37$, $p < .01$, $\eta^2 = .03$, so that Indian interviewees provided longer answer than Swiss interviewees ($p < .01$). But there was neither an effect of the interview medium, $F(2, 283) = 1.07$, $p = .35$, $\eta^2 = .01$, nor

an interaction effect, $F(2, 283) = 2.55$, $p = .08$, $\eta^2 = .02$ on interviewees' response length. Therefore, to investigate effects of the interview medium and culture on interviewees' production of STAR narrative elements, we conducted three two-way 3 (Interview Medium) x 2 (Culture) ANOVAs, using word count of participants' responses as a covariate to control for verbosity.

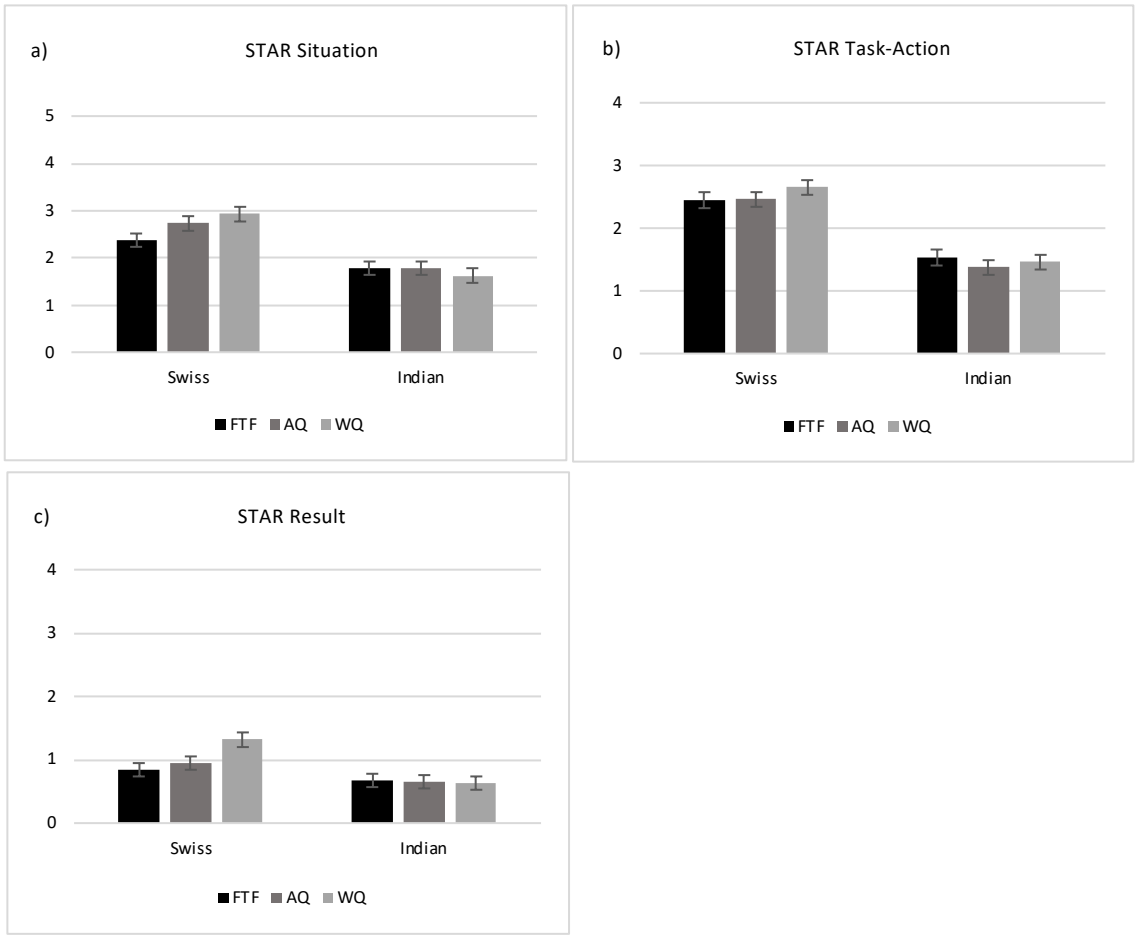
Results showed no main effect of the interview medium on the production of STAR Situation, $F(2, 282) = 1.47$, $p = .23$, $\eta^2 = .01$, on STAR Task-Action: $F(2, 282) = .94$, $p = .39$, $\eta^2 = .01$, or on STAR Result, $F(2, 282) = 2.67$, $p = .07$, $\eta^2 = .02$.

Results indicated a main effect of culture on all STAR elements (STAR Situation: $F(1, 282) = 81.07$, $p < .01$, $\eta^2 = .22$; STAR Task-Action: $F(1, 282) = 153.33$, $p < .01$, $\eta^2 = .35$; STAR Result: $F(1, 282) = 21.12$, $p < .01$, $\eta^2 = .07$). Pairwise comparisons showed that Swiss interviewees mentioned more of all STAR elements compared to their Indian counterparts ($p_s < .01$).

There was a significant interaction effect of the interview medium with culture on STAR Situation, $F(2, 282) = 3.84$, $p = .02$, $\eta^2 = .03$, and on STAR Result, $F(2, 282) = 3.54$, $p = .03$, $\eta^2 = .02$, but no significant interaction effect on STAR Task-Action, $F(2, 282) = .89$, $p = .41$, $\eta^2 = .01$. Concerning STAR Situation, pairwise comparisons showed that Swiss interviewees mentioned more situational elements in the AQ ($p = .05$) and WQ ($p < .01$) conditions than in the FTF condition. However, there was no difference between AQ and WQ conditions for Swiss interviewees ($p = .27$). No significant differences were observed across conditions for Indian interviewees ($p_s > .37$). As for STAR Result, pairwise comparisons indicated that Swiss interviewees mentioned more result-related element in the WQ condition than in the AQ ($p < .01$) and FTF ($p < .01$) conditions. But there was no significant difference between FTF and AQ condition for Swiss interviewees ($p = .45$). Again, no significant differences were found across conditions for Indian interviewees ($p_s > .78$). See Figure 1.

Figure 1.

Effect of Interview Medium and Culture on STAR Narrative Elements.



Note. Estimated means based on the Two-Way ANOVA results testing for the effect of interview medium and culture on STAR elements. FTF = face-to-face interviews; AQ = AVIs with avatar questions; WQ = AVIs with written questions.

Effect of Interview Medium and Culture on Disfluencies

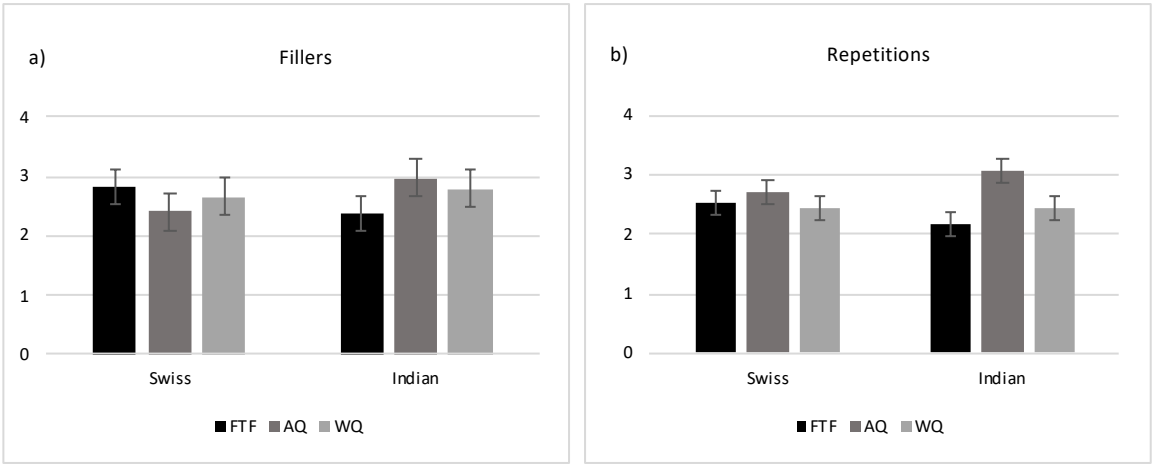
To investigate effects of the interview medium and culture on interviewees' disfluencies, we ran two two-way 3 (interview medium) x 2 (culture) ANOVAs. Results showed a significant effect of the interview medium on repetitions, $F(2, 283) = 4.70$, $p = .01$, $\eta^2 = .03$, but no significant effect on fillers, $F(2, 283) = .13$, $p = .88$, $\eta^2 = .00$. Concerning interviewees' repetitions, pairwise comparisons indicated that interviewees produced more

repetition in AQ condition than in FTF ($p < .01$) and WQ ($p = .02$) conditions. But there was no difference between FTF and WQ conditions ($p = .65$).

There was no main effect of culture on either fillers, $F(1, 283) = .17$, $p = .68$, $\eta^2 = .00$, or repetitions, $F(1, 283) = .00$, $p = .98$, $\eta^2 = .00$, and no interaction effect on either of the dependent variables (fillers: $F(2,283) = 1.76$, $p = .17$, $\eta^2 = .01$; repetitions: $F(2,283) = 1.70$, $p = .19$, $\eta^2 = .01$). See Figure 2.

Figure 2

Effect of Interview Medium and Culture on Disfluencies.



Notes. Estimated means based on the Two-Way ANOVA results testing for the effect of interview medium and culture on evaluations of participants. FTF = face-to-face interviews; AQ = AVIs with avatar questions; WQ = AVIs with written questions.

Effect of Interview Medium and Culture on Raters' Evaluations

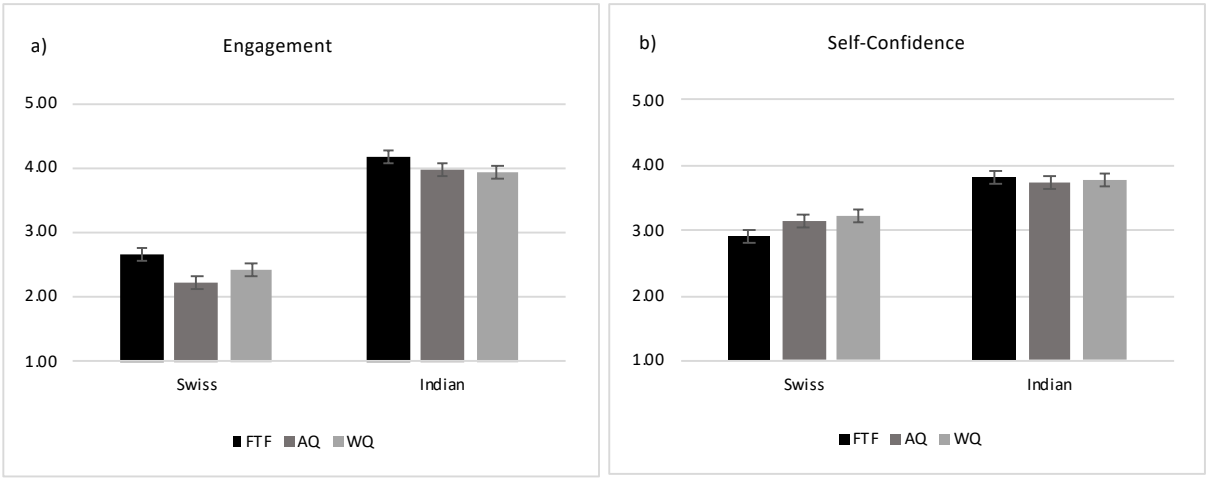
To investigate the effects of the interview medium and culture on evaluations of interviewees, we ran two two-way 3 (interview medium) x 2 (culture) ANOVAs. Results showed a main effect of the interview medium on perceived engagement, $F(2,283) = 4.59$, $p = .01$, $\eta^2 = .03$. But there was no significant main effect of the interview medium on perceived self-confidence, $F(2,283) = 1.22$, $p = .30$, $\eta^2 = .01$. Pairwise comparisons showed that interviewees were perceived as more engaged in the FTF than in the AQ ($p < .01$) and the WQ ($p = .03$) conditions. But perceived engagement did not differ between the AQ and WQ conditions ($p = .47$).

Results also indicated a main effect of culture on perceived engagement, $F(1,283) = 323.36$, $p < .01$, $\eta^2 = .53$, and perceived self-confidence, $F(1,283) = 79.36$, $p < .01$, $\eta 2 = .22$. Indian participants were perceived as more engaged ($p < .01$) and self-confident ($p < .01$) than Swiss participants.

Finally, there was no significant interview medium x culture interaction for perceived engagement, $F(2,283) = .86$, $p = .42$, $\eta^2 = .01$, or perceived self-confidence, $F(2,283) = 2.07$, $p = .13$, $\eta^2 = .01$. See Figure 3.

Figure 3.

Effect of Interview Medium and Culture on Evaluations of Participants.



Notes. Estimated means based on the Two-Way ANOVA results testing for the effect of interview medium and culture on evaluations of participants. FTF = face-to-face interviews; AQ = AVIs with avatar questions; WQ = AVIs with written questions.

Discussion

Main Findings and Theoretical Implications

Given the widespread adoption of AVIs in personnel selection, it is important to better understand how such new interview modalities can impact applicants' behaviors and performance in the interview process. Despite their flexibility, AVIs lack the interactivity of FTF interviews (Basch et al., 2020), and we know little about potential cross-cultural differences and associated biases (Arseneault & Roulin, 2023). Nevertheless, research suggests that applicants' performance in AVIs may, in some cases, be similar to their performance in

FTF interviews across cultures (Kleinlogel et al., 2023). Still, we currently know little about specific comparisons of different AVI designs with FTF interviews. The present research had four objectives: To investigate the effect of interview medium on (1) applicant storytelling responses to past-behavior questions and (2) disfluencies in behavioral interviews; their effect on (3) raters' evaluations of applicants' perceived engagement and self-confidence; and (4) cross-cultural differences in those variables.

Considering the impact of interview medium on storytelling, results indicated that interviewees consistently produced more stories in AVIs (both with avatar-based and written questions) than in FTF interviews. But the medium did not affect the quantity of narrative elements (STAR) provided when controlling for interviewees' verbosity (which was higher in Indian participants). These results may be because AVIs are less rich, and interviewees have no access to cues from recruiters (except for cues like the time available to respond). Therefore, they may be less hesitant to delay the initiation of their response while searching for a suitable episode to narrate, contrary to FTF interviews where applicants are under time pressure to initiate their response (Brosy et al., 2016) because of the recruiter sitting opposite them. Also, applicants in AVIs with WQ may benefit from the question being displayed on-screen to guide their answer. Interestingly, Swiss interviewees mentioned more situational elements in their answers to AVIs with WQ than in FTF interviews, and more result-related elements in AVIs with WQ compared to both FTF interviews and avatar-based AVIs. Interviewees may have felt that situation and result-related elements were more necessary to contextualize their stories by establishing common ground (Clark, 1996) in the interview medium where another human was not present (vs. in the FTF interviews conducted by humans).

Regarding the effect of culture on storytelling across interview media, Swiss interviewees produced storytelling responses more often than Indian interviewees. Furthermore, we observed cultural differences in the mention of STAR elements, thus

indicating differences in how narrative responses are constructed across cultures. These differences may stem from cross-cultural variations of social norms (Liu et al., 2022). Interviewees' expectations and perceptions of job interview questions can also differ depending on their cultural norms. Indeed, typical job interview questions vary across different cultures. Some cultures (e.g., Taiwan) ask more commonly questions about personal values, opinions and beliefs than others (e.g., Russia) (Posthuma et al., 2014). In this way, Indian interviewees may have been less familiar with past-behavior questions, and thus less aware of the appropriateness of producing stories in response to such questions.

Regarding the impact of interview medium and culture on interviewees' disfluencies in behavioral interviews, repetitions were higher in AQ interviews compared to FTF and WQ interviews, but they did not differ across cultures. Additionally, there were no differences in interviewees' use of fillers between FTF, AQ and WQ interviews, nor between cultures. Brosy et al. (2016) suggested that applicants' disfluencies may stem from the choice between delaying their responses to offer a higher-quality answer or responding quickly with a less relevant one, rather than keeping recruiters waiting. However, subsequent research (Brosy et al., 2020) indicated that disfluencies in responses are not solely attributable to the two-way interaction, but are most likely influenced by the challenges applicants encounter in finding an appropriate answer. Our study provided additional evidence that disfluencies may result from the search for a perfect answer rather than the interaction with the recruiter (Brosy et al., 2020). Also, the higher repetition rate in AQ interviews may reflect the unfamiliar nature of the avatar-based interactions. Interviewees may perceive this type of interview creepy (Langer & König, 2018), potentially impacting their speech fluency. Moreover, our results suggest that disfluent speech due to the search for the perfect answer is a challenge for applicants in both cultures, even though representations of appropriate behaviors may vary across cultures (Liu et al., 2022).

In examining the effects of the interview medium and culture on raters' evaluations, we found that interviewees were perceived as more engaged in FTF interviews compared to AVIs, with no differences between AVIs with AQ and WQ. A possible explanation stems from the greater co-presence and interactivity inherent to FTF interviews (Basch et al., 2020). This dynamic interaction may not only foster rapport-building, but also enhance overall engagement. Furthermore, our results suggest that applicants from different cultural backgrounds may be assessed differently: Indian interviewees were perceived as more engaged and self-confident than their Swiss counterparts. Nevertheless, Kleinlogel et al. (2023) found that Swiss and Indian participants were rated similarly in terms of perceived performance and stress. Also, it is important to note that in this study, the raters evaluating self-confidence and engagement of Indian participants were Swiss. We thus cannot exclude the possibility of cultural biases (Arseneault & Roulin, 2023).

Practical Implications

The findings have practical implications for job applicants and hiring organizations. First, they underscore the importance of considering the interview medium and its effect on applicants. They particularly reinforce prior advice (Langer et al., 2017; Melchers et al., 2021) to use the same interview medium across all applicants within the same selection process. Second, organizations may not necessarily need to invest in costly developments of avatar-based AVIs since they were associated with results similar to (cheaper) AVIs with written questions displayed on-screen in terms of applicant responses or performance (Kleinlogel et al., 2023) and raters' evaluations. This is particularly important because avatar-based AVIs could constitute a yet unfamiliar format, and therefore be perceived as creepier by applicants (Langer & König, 2018). Lastly, organizations engaging in international recruitment and selection should consider culture both when designing interview questions and evaluating applicants. This is crucial because variations in applicants' responses may not only be due to

individual differences, but also to cultural differences. For example, while past-behavior questions are a best practice, we should consider that cultural differences can sometimes hinder applicants in producing stories and thus responding appropriately.

Limitations and Future Research

The present study has limitations that future research should consider. The first limitation concerns the sample of male participants aged 18 to 49. Results may not generalize to females or to individuals above 50 who might respond differently to technology.

A second limitation pertains to the fact that data were collected through mock job interviews in lab experiments. Results may not generalize in a straightforward manner to high-stakes selection interviews, despite our efforts to replicate real-world conditions (e.g., let interviewees select a job relevant to them). Nonetheless, conducting this study in a lab setting provided rigorous control and thus stronger causal evidence.

Furthermore, the observation of cross-cultural differences in applicant responses to past-behavior questions is specific to Swiss and Indian applicants. Also, reliability for interviewees' storytelling was lower for the Swiss data than for the Indian data. Additionally, the raters responsible for evaluating Indian participants' perceived engagement and self-confidence were of Swiss origin, which may have introduced cultural bias. Therefore, findings should be interpreted with caution and cannot be generalized to other cultural contexts. Nonetheless, the identified cultural distinctions shed light on potential differences in behavioral selection interviews across various cultures. These limitations underscore the need for future research with more diverse and representative samples, conducted in real-world settings, and involving broader cultural comparisons to enhance the robustness and applicability of the findings.

Conclusion

The present study contributes to understanding similarities and differences between AVIs and traditional in-person interviews. Through a systematic examination of applicant responses in behavioral in FTF interviews and different AVI settings and across culture, we found that applicants' speech fluency is comparable between interview media and different cultures. However, we observed different patterns of storytelling and raters' evaluation across interview media and cultures.

References

Arseneault, R., & Roulin, N. (2021). A theoretical model of cross-cultural impression management in employment interviews. *International Journal of Selection and Assessment*, *29*(3-4), 352-366. https://doi.org/10.1111/ijsa.12348

Arseneault, R., & Roulin, N. (2023). Examining discrimination in asynchronous video interviews: Does cultural distance based on country-of-origin matter? *Applied Psychology*, 1–30. https://doi.org/10.1111/apps.12471

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. *Journal of Business and Psychology*, *29*(4), 593-604. https://doi.org/10.1007/s10869-014-9350-0

Basch, J. M., & Melchers, K. G. (2019). Fair and flexible?! Explanations can improve applicant reactions toward asynchronous video interviews. *Personnel Assessment and Decisions*, *5*(3), 2. https://doi.org/10.25035/pad.2019.03.002

Basch, J. M., Melchers, K. G., Kegelmann, J., & Lieb, L. (2020). Smile for the camera! The role of social presence and impression management in perceptions of technology-mediated interviews. *Journal of Managerial Psychology 35*(4), 285-299. https://doi.org/10.1108/JMP-09-2018-0398

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology 79*(6), 941-952.

Bencharit, L. Z., Ho, Y. W., Fung, H. H., Yeung, D. Y., Stephens, N. M., Romero-Canyas, R., & Tsai, J. L. (2019). Should job applicants be excited or calm? The role of culture and ideal affect in employment settings. *Emotion*, *19*(3), 377-401. https://doi.org/10.1037/emo0000444

Brennan, S. E., & Williams, M. (1995). The feeling of another′ s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383-398. https://doi.org/10.1006/jmla.1995.1017

Brosy, J., Bangerter, A., & Mayor, E. (2016). Disfluent responses to job interview questions and what they entail. *Discourse Processes*, *53*(5-6), 371-391. https://doi.org/10.1080/0163853X.2016.1150769

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. *European Journal of Work and Organizational Psychology*, *29*(3), 330-343. https://doi.org/10.1080/1359432X.2019.1704265

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human Communication Research*, *17*(1), 140-169. https://doi.org/10.1111/j.1468-2958.1990.tb00229.x

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655-702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589-602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, *32*(5), 554-571. https://doi.org/10.1287/mnsc.32.5.554

Day, A. L., & Carroll, S. A. (2003). Situational and patterned behavior description interviews: A comparison of their validity, correlates, and perceived fairness. *Human Performance*, *16*(1), 25-47. https://doi.org/10.1207/S15327043HUP1601_2

DeGroot, T., & Gooty, J. (2009). Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology*, *24*(2), 179-192. https://doi.org/10.1007/s10869-009-9098-0

DeGroot, T., & Motowidlo, S. J. (1999). Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology*, *84*(6), 986-993. https://doi.org/10.1037/0021-9010.84.6.986

Feiler, A. R., & Powell, D. M. (2016). Behavioral expression of job interview anxiety. *Journal of Business and Psychology*, *31*(1), 155-171. https://doi.org/10.1007/s10869-015-9403-z

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Finnerty, A. N., Muralidhar, S., Nguyen, L. S., Pianesi, F., & Gatica-Perez, D. (2016). Stressful first impressions in job interviews. Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan.

Forbes, R. J., & Jackson, P. R. (1980). Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, *53*(1), 65-72. https://doi.org/0.1111/j.2044-8325.1980.tb00007.x

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*(6), 709-738. https://doi.org/10.1006/jmla.1995.1032

Frauendorfer, D., & Schmid Mast, M. (2015). The impact of nonverbal behavior in the job interview. In A. Kostić & D. Chadee (Eds.), *The Social Psychology of Nonverbal*

Communication (pp. 220-247). London: Palgrave Macmillan. https://doi.org/10.1057/9781137345868_11

Gifford, R., Ng, C. F., & Wilkinson, M. (1985). Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, *70*(4), 729-736. https://doi.org/10.1037/0021-9010.70.4.729

Hartwell, C. J., Johnson, C. D., & Posthuma, R. A. (2019). Are we asking the right questions? Predictive validity comparison of four structured interview question types. *Journal of Business Research*, *100*, 122-129. https://doi.org/10.1016/j.jbusres.2019.03.026

Hirevue. (2022). *HireVue milestone and major company growth*. Retrieved 12.07.2023 from https://www.hirevue.com/press-release/hirevue-supercharged-for-growth-hitting-33-million-interviews-milestone-and-appointing-chief-growth-officer-patrick-morrissey-chief-marketing-officer-amanda-hahn

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5). Sage Publications.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Culture and organizations. Software of the mind: Intercultural cooperation and its importance for survival*. NewYork, NY : McGraw Hill. https://doi.org/10.1080/00208825.1980

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897-913. https://doi.org/10.1037/0021-9010.86.5.897

Imada, A. S., & Hakel, M. D. (1977). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology*, *62*(3), 295-300. https://doi.org/10.1037/0021-9010.62.3.295

Jan, J., Alshare, K. A., & Lane, P. L. (2022). Hofstede's cultural dimensions in technology acceptance models: a meta-analysis. *Universal Access in the Information Society*, 1-25. https://doi.org/10.1007/s10209-022-00930-7

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology*, *67*(5), 577-580. https://doi.org/10.1037/0021-9010.67.5.577

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145-167. https://doi.org/10.1080/01690960802586188

Kleinlogel, E. P., Schmid Mast, M., Jayagopi, D. B., Shubham, K., & Butera, A. (2023). "The interviewer is a machine!" Investigating the effects of conventional and technology-mediated interview methods on interviewee reactions and behavior. *International Journal of Selection and Assessment*, *31*, 403-419. https://doi.org/10.1111/ijsa.12433

Langer, M., & König, C. J. (2018). Introducing and Testing the Creepiness of Situation Scale (CRoSS) [Original Research]. *Frontiers in Psychology*, *9:2220*. https://doi.org/10.3389/fpsyg.2018.02220

Langer, M., König, C. J., & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. *International Journal of Selection and Assessment*, *25*(4), 371-382. https://doi.org/10.1111/ijsa.12191

Leigh, T. W., & Summers, J. O. (2002). An initial evaluation of industrial buyers' impressions of salespersons' nonverbal cues. *Journal of Personal Selling & Sales Management*, *22*(1), 41-53. https://doi.org/10.1080/08853134.2002.10754292

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

Liu, R. W., Lapinski, M. K., Kerr, J. M., Zhao, J., Bum, T., & Lu, Z. (2022). Culture and Social Norms: Development and Application of a Model for Culturally Contextualized Communication Measurement (MC3M). *Frontiers in Communication*, *6:770513*. https://doi.org/10.3389/fcomm.2021.770513

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), 1-15. https://doi.org/10.1016/j.hrmr.2020.100789

Mandelbaum, J. (2013). Storytelling in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 492-507). Chichester, UK: John Wiley & Sons.

Manroop, L., Boekhorst, J. A., & Harrison, J. A. (2013). The influence of cross-cultural differences on job interview selection decisions. *The International Journal of Human Resource Management*, *24*(18), 3512-3533. https://doi.org/10.1080/09585192.2013.777675

Melchers, K. G., Petrig, A., Basch, J. M., & Sauer, J. (2021). A comparison of conventional and technology-mediated selection interviews with regard to interviewees' performance, perceptions, strain, and anxiety. *Frontiers in Psychology*, *11:603632*. https://doi.org/10.3389/fpsyg.2020.603632

Mino, M. (1996). The relative effects of content and vocal delivery during a simulated employment interview. *Communication Research Reports*, *13*(2), 225-238. https://doi.org/10.1080/08824099609362090

Motowidlo, S. J. (1999). Asking about past behavior versus hypothetical behavior. In R. W. Eder & H. M. M. (Eds.), *The employment interview handbook* (pp. 179-190). Sage Publications.

Nguyen, L. S., & Gatica-Perez, D. (2015). I would hire you in a minute: Thin slices of nonverbal behavior in job interviews. Proceedings of the 2015 ACM on international conference on multimodal interaction, Seattle, Washington, USA.

Parsons, C. K., & Liden, R. C. (1984). Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology*, *69*(4), 557-568. https://doi.org/10.1037/0021-9010.69.4.557

Posthuma, R. A., Levashina, J., Lievens, F., Schollaert, E., Tsai, W.-C., Wagstaff, M. F., & Campion, M. A. (2014). Comparing employment interviews in Latin America with other countries. *Journal of Business Research*, *67*(5), 943-951. https://doi.org/10.1016/j.jbusres.2013.07.014

Potosky, D. (2008). A conceptual framework for the role of the administration medium in the personnel assessment process. *Academy of Management Review*, *33*(3), 629-648. https://doi.org/10.5465/amr.2008.32465704

Ralston, S. M., Kirkwood, W. G., & Burant, P. A. (2003). Helping interviewees tell their stories. *Business Communication Quarterly*, *66*(3), 8-22. https://doi.org/10.1177/108056990306600303

Roulin, N., Bangerter, A., & Wüthrich, U. (2012). *Réussir l'entretien d'embauche comportemental: La méthode pour identifier et sélectionner les futurs employés performants*. Bruxelles: De Boeck Professionals.

Sandal, G. M., van de Vijver, F., Bye, H. H., Sam, D. L., Amponsah, B., Cakar, N., Franke, G. H., Ismail, R., Kjellsen, K., & Kosic, A. (2014). Intended self-presentation tactics in job interviews: A 10-country study. *Journal of Cross-Cultural Psychology*, *45*(6), 939-958. https://doi.org/10.1177/00220221145323

Stenström, A. (2011). Pauses and hesitations. In G. Andersen & K. Aijmer (Eds.), *Pragmatics of society* (pp. 537-568). De Gruyter Mouton.

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, *80*(5), 587-606. https://doi.org/10.1037/0021-9010.80.5.587

Tessler, R., & Sushelsky, L. (1978). Effects of eye contact and social status on the perception of a job applicant in an employment interviewing situation. *Journal of Vocational Behavior*, *13*(3), 338-347. https://doi.org/10.1016/0001-8791(78)90060-X

Triandis, H. C. (1982). Review of culture's consequences: International differences in work-related values. *Human Organization*, *41*(1), 86-90.

Van Iddekinge, C. H., Raymark, P. H., Roth, P. L., & Payne, H. S. (2006). Comparing the psychometric characteristics of ratings of face-to-face and videotaped structured interviews. *International Journal of Selection and Assessment*, *14*(4), 347-359. https://doi.org/10.1111/j.1468-2389.2006.00356.x

Vijay, R. S., Shubham, K., Renier, L. A., Kleinlogel, E. P., Mast, M. S., & Jayagopi, D. B. (2021). An Opportunity to Investigate the Role of Specific Nonverbal Cues and First Impression in Interviews using Deepfake Based Controlled Video Generation. Companion Publication of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada.

III Appendix 3: Study 3

Germanier, E., Bangerter, A., Orji, K., Schmid Mast, M., Renier, L. A., He, M., & Garner, P. N. (2024). *Responses to Past-Behavior Questions in Face-to-Face and Asynchronous Video Interviews: Storytelling, Interview Performance and Criterion Validity* [Manuscript submitted for publication]. Institute of Work and Organizational Psychology, University of Neuchâtel

Status: The manuscript is currently under revision. The version in this thesis is an intermediate between the submitted manuscript and the fully revised one.

Responses to Past-Behavior Questions in Face-to-Face and Asynchronous Video Interviews: Storytelling, Interview Performance and Criterion Validity

Elisabeth Germanier[1], Adrian Bangerter[1], Koralie Orji[1], Laetitia A. Renier[2], Marianne Schmid Mast[2], Mutian He[3], Philip N. Garner[3]

[1] Institute of Work and Organizational Psychology, University of Neuchâtel, Switzerland

[2] Department of Organizational Behavior, University of Lausanne, Switzerland

[3] Idiap Research Institute, Martigny, Switzerland

Author Note

Elisabeth Germanier https://orcid.org/0009-0006-1220-8601,

Adrian Bangerter https://orcid.org/0000-0001-6989-8654,

Koralie Orji https://orcid.org/0009-0004-2879-8082,

Laetitia A. Renier https://orcid.org/0000-0002-9686-1686,

Marianne Schmid Mast https://orcid.org/0000-0001-5510-610X,

Mutian He https://orcid.org/0000-0002-8939-4207,

Philip N. Garner https://orcid.org/0000-0002-0814-1348

The authors report no conflict of interest.

This work was supported by the Swiss National Science Foundation (SNSF) under Grant 10521C_197479.

Correspondence concerning this article should be addressed to Elisabeth Germanier, Institute of Work and Organizational Psychology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. E-Mail: elisabeth.germanier@unine.ch.

Word count: 7404

Abstract

Past-behavior questions are valid predictors of job performance and have high criterion validity in face-to-face (FTF) interviews. However, their criterion validity in asynchronous video interviews (AVIs) is under-researched. Our experiment investigated the effect of interview medium (FTF versus AVI) on participants' responses to past-behavior questions, interview performance, and the criterion validity of past-behavior questions. Participants (n = 229) completed a standardized work sample as a measure of exercise performance followed by a mock interview in an AVI or FTF setting, in which one past-behavior question targeted that work sample. Participants' responses and interview performance were similar across interview media, but they described more values and opinions in AVIs. Participants' overall interview performance in AVIs and FTF settings were valid predictors of exercise performance, and their performance at the past-behavior question targeting the work sample also.

Keywords: Asynchronous video interview, past-behavior questions, storytelling, performance, criterion validity.

Responses to Past-Behavior Questions in Face-to-Face and Asynchronous Video Interviews: Storytelling, Interview Performance and Criterion Validity

Introduction

Behavioral questions constitute a best practice in job interviews. According to a recent survey, 75% of recruiters assess applicants' characteristics using behavioral questions (LinkedIn, 2019), often focusing on applicants' past behaviors (Janz, 1989). Past-behavior questions (e.g., "Can you describe a situation where you had to deal with a conflict in your team?") invite applicants to recount their actions in past work-related situations. Recruiters use their answers to assess applicants' mastery of competencies or other personal characteristics (Roulin et al., 2012; Turner, 2004). Past-behavior questions constitute invitations to applicants to produce a story (Bangerter et al., 2014; Brosy et al., 2020; Ralston et al., 2003; Stevens & Kristof, 1995). These questions are valid predictors of job performance (Huffcutt et al., 2004; Janz, 1989). However, their construct validity is low (Van Iddekinge et al., 2004): It remains unclear whether they target the specific competencies they are supposed to measure or other characteristics such as communication skills (Burroughs & White, 1996).

Job interviews are undergoing technological innovations, evidenced by the rising popularity of computer-mediated interviews such as asynchronous video interviews (AVIs; Ciphr, 2023). In AVIs, applicants connect to a web-based platform where they answer questions written on-screen or asked by an avatar and record their responses using their computer webcam (Lukacik et al., 2022). While AVIs bring important advantages (e.g., scalability, scheduling flexibility), they fundamentally alter the classical interview dynamics between applicants and recruiters (Basch et al., 2020). In the absence of real-time interactions with recruiters, applicants' responses to past-behavior questions and performance in AVIs may

differ from FTF interviews. There is very little research addressing this issue. In an initial study, interview performance was similar between the two interview media (Kleinlogel et al., 2023). Moreover, potential differences in response content across media may jeopardize the criterion validity of past-behavior questions for predicting job performance. A recent study found that past-behavior questions in AVIs have high criterion validity (Liff et al., 2024). However, the job performance measure was based on supervisors' and customers' evaluations that may hold biases. Moreover, it is unknown whether the criterion validity of past-behavior questions in AVIs is comparable to FTF interviews. More data is needed to gain a comprehensive understanding of these issues.

This experimental study compares applicants' responses to past-behavior questions, interview performance, and criterion validity of past-behavior questions in AVIs and FTF interviews. Participants completed a standardized work sample (role-play), from which we derive a measure of the exercise performance. A week later, they completed either an AVI or a FTF interview featuring past-behavior questions, including one question targeting the work sample. This study design allows (1) analyzing participants' response types (e.g., storytelling), (2) evaluating their interview performance using behaviorally-anchored rating scales (BARS) and (3) assessing the link between interview and exercise performance at the work sample (i.e., criterion validity) in both media.

This study fills several gaps for both research and practice. It complements initial research about criterion validity of past-behavior questions in AVIs (Liff et al., 2024) by comparing it to FTF interviews and using a work sample as a measure of exercise performance. Furthermore, it offers a detailed comprehension of applicants' response content across FTF interviews and AVIs. The study also provides further evidence on applicants' interview performance across interview media (Kleinlogel et al., 2023). From a practical perspective, our findings may help recruiters use past-behavior questions effectively in various interview

contexts. Moreover, insights into how applicants respond to past-behavior questions in AVIs enable organizations to design interviews to ensure that applicants can perform at their best.

Past-Behavior Questions

Past-behavior questions constitute a best practice for assessing applicant characteristics and predicting future workplace performance (Campion et al., 1997). They aim to elicit accounts of applicants' behaviors in a past work-related situation, under the premise that those behaviors are likely to be the same in the future, and thus predictive of future job performance (Janz, 1989). To evaluate applicants' responses, recruiters are recommended to use behaviorally anchored rating scales (BARS) to mitigate bias (Smith & Kendall, 1963). Past-behavior questions have high criterion validity (e.g., Motowidlo et al., 1992; Huffcutt et al., 2004; Taylor & Small, 2010; Hartwell et al., 2019). However, the constructs measured by past-behavior questions remain unclear. While one study reports a strong convergent correlation between recruiters' and job supervisors' evaluations (Motowidlo et al., 1992), others suggest lower construct validity. This means that interview ratings might not capture the constructs that are targeted, but rather more general characteristics (Van Iddekinge et al., 2004) like job knowledge, past experiences, communication skills, or general mental aptitude (e.g., Borroughs & White, 1996; Conway & Peneno, 1999; Salgado & Moscoso, 2002; Huffcutt & Murphy, 2023).

Besides psychometric issues, how applicants respond to past-behavior questions remains poorly understood. Optimal responses involve describing a specific problematic work-related situation, one's actions in the situation and how these led to a resolution of the problem (Ralston et al., 2003; Stevens & Kristof, 1995). Applicants should thus engage in *storytelling* with the help of recruiters playing the role of active listeners (Bavelas et al., 2000; Mandelbaum, 2013). Advice literature suggests applicants should structure their stories using the STAR method, that is, describe the initial situation (S), their tasks (T) and actions (A), and

obtained results (R) (Kessler, 2006). Applicants who produce storytelling responses tend to receive better hiring recommendations (Bangerter et al., 2014), perhaps because their responses create more distinctive and therefore more memorable impressions (Stevens & Kristof, 1995).

However, applicants may have difficulties in effectively generating compelling stories on the spot (Bangerter et al., 2014). They may struggle to recall an appropriate past situation, organize their thoughts, and deliver a coherent answer rapidly after the recruiters' question (Brosy et al., 2016; Brosy et al., 2020). Thus, they may retrieve instances of typical, rather than maximal performance, from memory (Huffcutt et al., 2024). Those episodes may be recounted as generic descriptions of typical situations (i.e., *pseudo-stories*). If pressed for time to respond, they may resort to decontextualized assertions such as self-descriptions (*I am someone who likes a job well-done*) and expressions of their values and opinions (*I think it's important to be fair in all situations*) (Bangerter et al., 2014). Further, when applicants do engage in storytelling, they tend to focus on describing the initial situation rather than detailing their actions and results (Bangerter et al., 2014). However, recruiters can play an important role in optimizing question phrasing to prime applicants to recount their best performances (Huffcutt et al., 2024), and in facilitating more complete storytelling by using probe questions (Brosy et al., 2020).

Responding to Past-Behavior Questions in AVIs

Research on the validity of past-behavior questions and on applicants' responses and performance has predominantly focused on FTF interview settings. Yet, the emergence of computer-mediated job interviews may transform the dynamics of information exchange between recruiters and applicants. Recruiters can now resort to traditional in-person interviews or opt for an AVI via online platforms. In recent years, AVIs have gained interest within organizations due to their cost efficiency and flexibility (Mejia & Torres, 2018).

However, the absence of recruiters in AVIs deprives applicants of the naturally occurring conversational feedback that helps them gauge and adapt their responses on the fly. This feedback is crucial for storytelling. In traditional interaction settings, the audience co-creates the narration by continuously producing signals (e.g., backchannels) that help the narrator progress in the story (Mandelbaum, 2013). The audience thus actively contributes to the quality of storytelling. Indeed, experimental studies show that narrators tell their stories less effectively when audiences are distracted and unable to provide feedback (Bavelas et al., 2000). In AVIs, applicants' storytelling responses may be affected by recruiters' absence, thus differing from FTF interviews. Also, recruiters' absence may be even more impactful because they play an important role in probing applicants' responses (Brosy et al., 2020). How applicants' storytelling responses and other response types (e.g., pseudo-stories, decontextualized assertions) in AVIs compare to FTF interviews is unknown. For all these reasons, more research on the effects of interview medium on applicants' responses to past-behavior questions is necessary:

Research Question 1. What effect does the interview medium have on applicants' responses to past-behavior questions?

Applicants' interview performance varies across the different interview media. Applicants get lower performance ratings in videoconferences than in FTF interviews (Melchers et al., 2021) and lower performance ratings in videoconferences than in AVIs (Langer et al., 2017). But one study comparing FTF interviews with AVIs found that applicants' performance ratings were comparable across both media when using the same evaluation procedure and eliminating preparation time in the asynchronous condition (Kleinlogel et al., 2023). To provide additional results, further investigation is warranted:

Research Question 2. What effect does the interview medium have on applicants' interview performance?

Applicants' negative reactions toward AVIs may jeopardize the criterion validity of behavioral interviews. When applicants perceive less social presence in job interviews, they also perceive less opportunity to perform (Basch et al., 2020). This could affect their motivation to engage in the interview fully (Hausknecht et al., 2004). However, motivation to engage with a test is crucial for validity (Schmit & Ryan, 1992). Reduced motivation to perform in AVIs may lead recruiters to misinterpret the applicants' competencies and misevaluate their past job performance. Moreover, applicants experience more stress in AVIs than in FTF interviews (Kleinlogel et al., 2023), and higher interview anxiety can affect validity (Schneider et al., 2019).

There are two ways to assess the criterion validity of past-behavior questions. One way relies on asking for supervisors' evaluations as a measure of job performance (Hartwell et al., 2019; Motowidlo et al., 1992). Initial research suggests that behavioral questions in AVIs have good criterion validity in this respect (Liff et al., 2024). However, those evaluations may be biased by the employee's attributes (e.g., sex, ethnicity, attractiveness), liking, or even team size (Breuer et al., 2013; Lefkowitz, 2000). To tackle these evaluation biases, another way involves testing applicants' performance in a standardized work sample (Heimann et al., 2020; Krajewski et al., 2006), where performance can be compared under controlled conditions. As the validity of selection tools is central to the success of recruiting organizations, further investigation of past-behavior question criterion validity based on work samples in both interview media is warranted :

Research Question 3. What effect does the interview medium have on the criterion validity of past-behavior questions as measured in an experimental work sample?

Method

Sample

The sample size was determined before data collection based on a power analysis assuming a small to medium effect size, an alpha of .05 and a power of .95 with a repeated measure design and two experimental conditions, which resulted in a sample of 228 participants (Faul et al., 2009). Two hundred and fifty-four participants were recruited through the participant pool of a Swiss university after obtaining ethics clearance. Twenty-five participants were excluded for at least one of the following reasons: (1) they did not correctly follow instructions (e.g., completed questionnaires before completing the tasks; $n = 6$), (2) they knew one of the actresses involved in the role-play task ($n = 10$), (3) one of the actresses did not play according to the script (e.g., forgot to make a key statement; $n = 7$) or (4) videorecording failures ($n = 2$). This resulted in a final sample of $N = 229$ (61.1% women).

Participants' mean age was 21.55 years ($SD = 3.09$). Regarding their education level, 52% had a high school diploma, 38.4% had a bachelor's degree and 8.3% had a master's degree, 1.3% reported other degrees. On average, participants had a prior experience of 1.92 job interviews ($SD = 2.78$), and 19.2% had previous work experience. In compensation for the 1.5 hours dedicated to our study, participants received a fixed sum of 40 CHF (about 45 USD) and an additional performance-related bonus from 0 CHF to 40 CHF to increase motivation.

Design

We conducted a between-subject (FTF versus AVI) experiment in two sessions. In a first session, participants performed a standardized work sample from Richter et al. (2016) involving two successive role-plays where they played the role of a manager whose task was to lay off two employees (played by actresses). From these layoffs, we derived a measure of exercise performance. We had participants complete the layoff task twice because it was a stressful task in which participants were unlikely to have prior experience. We wanted them to

have the occasion to develop experience in how to handle the layoffs between the first and second time.

In a second session about one week later, they completed a mock interview in the FTF or AVI setting. Three past-behavior questions measured different competencies, including one specifically asking them to recount one of the previous week's layoffs. This design allows us to link the performance at the specific past-behavior question about the layoffs directly with the exercise performance at the layoffs and thus to compute the specific criterion validity of that past-behavior question. More generally, we can also link the overall interview performance (i.e., the three past-behavior questions) with exercise performance and thus compute overall interview criterion validity. Because participants' responses were video and audio-recorded and transcribed, our study further enables a detailed content analysis of participants' responses between FTF interviews and AVIs.

Procedure

The study comprised two sessions. In the first session, participants came to the lab to complete two role-plays. They read and signed a consent form and completed the HEXACO-PI-R (2004) personality questionnaire (data not reported here). They then completed the work sample (i.e., two layoffs). After each layoff, they self-reported their performance (data not reported here).

In the second session, about one week later, participants returned for the mock interview. They were randomly assigned to complete an interview in either the FTF or the AVI conditions. Finally, participants filled out the Honest Interview Impression Management questionnaire from Bourdage et al. (2018) (data not reported here).

Participants were video-recorded during the two layoff meetings and the job interview with front and side cameras, along with audio captured by a microphone. Also, AVIs were recorded through a computer webcam. For coding purposes, the interactions (i.e., the layoffs

and the job interview) were transcribed based on the audio-recording. In the few cases where audio-recordings were defective, we used the video-recording.

Materials and Tasks

Work Sample

To prepare for the layoffs, participants first read the instructions and the given scenario. The scenario was that their organization faced financial difficulties and had to reduce staff by 20% and outsource the customer service. As the service manager, they had to lay off two employees following specific criteria (i.e., age, seniority, and family obligations). Before each layoff, participants read the information about the employee they were to meet and personalize their preparation. Information was provided about employees' age, marital status, whether they had children and their age, some private information (e.g., mortgage, engaged in a divorce procedure), and their current job evaluations.

Participants prepared the first layoff for 10 minutes (with 5 minutes for instructions and 5 minutes for information about the first employee) and conducted the meeting for about 10 minutes. They then prepared the second layoff for 5 minutes (with information about the second employee) and conducted the meeting for about 10 minutes. The order in which each employee (actress) was laid off was counterbalanced.

During the layoffs, the actresses were trained to follow a script. In addition, they were trained to exhibit different behavioral styles, to confront participants with different situations, and thus force them to adapt their behavior between the first and second layoff. One actress was trained to enact aggressive reactions to participants' behavior (e.g., frustrated shouting). In contrast, another was trained to enact withdrawal reactions (e.g., remaining silent and avoiding the participant's gaze for long periods).

Job Interview

The experimenter briefed participants that the job interview was for a manager position in a company selling computers. Immediately after, participants completed the job interview. Both FTF interviews and AVIs featured the same four questions. The first question asked applicants to briefly present themselves (i.e., **"**Could you please introduce yourself and summarize your background in a few words?")**.** The second and third questions were past-behavior questions about managerial competencies and their order was counterbalanced (i.e., multitasking: "Can you describe a situation where you were responsible for completing several tasks in a short time? How did you organize yourself to manage this situation?"; team management: "Can you describe a situation where one of your colleagues did not complete the tasks requested within the time limit? How did you deal with this situation?"). The last past-behavior question was about the previous week's layoffs (i.e., "Can you tell us about a recent situation in which you had to give bad news to someone?").

In the FTF condition, participants answered interview questions asked by a human recruiter (an I/O psychology graduate student trained to follow the job interview script). The recruiter first invited participants to sit down, thanked them for coming to the job interview, explained the job interview procedure, and then asked the interview questions before concluding the interview. For the last question, the recruiter could specify the question if participants did not refer to one of the previous week's layoffs ("Could you please tell me about one of the two layoffs in particular?"). Throughout the interview, the recruiter could briefly answer any participants' question, produce backchannels (mainly *mhm*), or use words like "indeed" to demonstrate active listening.

In the AVI condition, the experimenter opened the same AVI platform as in Roulin et al's (2023) study. Participants read instructions on how to use the platform. Then, the experimenter ensured the proper functioning and calibration of the computer webcam before

leaving the room. During the interview, participants read a job interview introduction, responded to interview questions written on-screen, and read a closing text, all mirroring the recruiter's in-person script. Participants had up to 20 seconds to read each interview question and up to 5 minutes to answer the question before clicking the "next" button. They could start answering the question before the end of the 20-second countdown by clicking on a "start recording" button. Also, they could stop responding to the question before the 5 minutes were up by clicking on a "stop recording" button.

Measures

Response Type

To code for response type in participants' responses, each transcribed response to a past-behavior question was segmented into utterances beforehand. Response type was coded by two of the co-authors into one of 6 categories adapted from Bangerter et al. (2014, p. 598): (1) story (i.e., "set of events related to a unique past episode, characterized by a unity of time or action"), (2) pseudo-story (i.e., "a description of a generic situation or recurrent set of similar situations, without unity of time or action."), (3) decontextualized assertions about one's opinion (i.e., "value/opinion"), (4) self-descriptions, (5) justifications of one's actions and (6) other utterances. Each coder coded half of the data. Reliability was established by double-coding 30 transcripts (Fleiss' $\kappa = .99$, $p < .01$). The occurrences of each response type were summed to calculate the total number of utterances falling into each category over all responses of a participant.

Exercise Performance

Exercise performance was assessed based on video recordings of the layoff interactions by two pairs of trained psychology graduate students using an adapted version of Richter et al.'s (2016) checklist for delivering bad news. Our adaptation featured 27 items, a sample item is "Stays objective and calm*"*. For each item, raters coded as "1" if the item was present;

otherwise, it was "0". All items were summed so that exercise performance scores ranged from 0 (= poor performance) to 27 (= excellent performance). To prevent a halo effect, we used two pairs of coders: one pair evaluated the first layoff, and the other pair evaluated the second layoff. This ensured that each coder rated a participant only once. Interrater agreement, calculated as the correlation between both ratings of a layoff, was high (both rs = .84, $p < .001$). We first averaged both ratings to create a score for each participant's performance for each layoff. Since both layoff performances significantly correlated ($r = .43$, $p < .001$), we then computed overall exercise performance by averaging scores across both layoffs.

Interview Performance

Interview performance was assessed based on interview video recordings by two trained graduate students using self-developed BARS with definitions and examples of behaviors (Smith & Kendall, 1963). Each participant's answer to a past-behavior question was rated from 1 (=not at all competent) to 5 (=fully competent). If the FTF recruiter asked a probe question about the layoff, raters only evaluated participants' responses before the probe question. Also, they were instructed to assess participants' performance based solely on explicitly described behaviors. This instruction led to low interview performance scores (M = 1.63, SD = .42). Interrater agreement was calculated as the correlation between both ratings of an interview and was high ($r = .81$, $p < .001$). To get a single measure for each participant's response, we initially computed three interview performance scores (one per response to a past-behavior question) by averaging the two raters' evaluations. Then, we determined the overall interview performance by averaging these three scores.

Results

Descriptive statistics by condition are presented in Table 1. Correlations for study variables are presented in Table 2. The analyses were performed using R 4.1.1 (R Core Team, 2021) and the following R packages: tidyverse 1.3.2 (Wickham et al., 2019) for data

manipulation and visualization, effectsize 0.8.6 (Ben-Shachar et al., 2020) to obtain effect sizes, moments 0.14.1 (Komsta & Novomestky, 2022) to analyze the data distribution, car 3.1.0 (Fox & Weisberg, 2019) to compute the analyses of variance. We used PROCESS for R 4.3.1 (Hayes, 2017) for moderation analyses.

Effect of Interview Medium on Applicants' Responses

Before testing the impact of the interview medium on response types, we checked whether it influenced participants' overall verbosity. Participants produced significantly more utterances in AVIs, (M = 105.71, SD = 45.09) than in FTF interviews (M = 62.42, SD = 30.77), $t(201.44) = 8.49$, $p < .001$, Cohen's d = 1.12, 95% CI [33.24, 53.34]. For subsequent analyses of participants' response types, we thus controlled for verbosity by computing ratios of response types. We divided the counts of each response type by the total number of utterances produced by the participant while answering past-behavior questions.

To investigate the impact of the interview medium on participants' response types (Research Question 1), we conducted a MANOVA. Results revealed a significant main effect of the interview medium on response type, $F(1,227) = 3.18$, $p = .005$, $\eta p^2 = .08$.

To determine which response types were influenced by the interview media, we further investigated the effect of the interview medium on each response types with six univariate ANOVAs. To avoid Type I errors due to multiple tests, we used the Bonferroni correction (corrected $\alpha = .008$). Results showed a significant effect of the interview medium on Values/Opinion, $F(1,227) = 16.59$, $p < .001$, $\eta p^2 = .07$, which were more frequent in AVIs (M = 22.32, SD = 13.48) than in FTF interviews (M = 15.42, SD = 11.69). However, there was no significant effect of interview medium on the production of stories ($F(1,227) = 2.54$, $p = .112$, $\eta p^2 = .01$), on the production of pseudo-stories ($F(1,227) = 0.21$, $p = .646$, $\eta p^2 = .00$), on self-

descriptions ($F(1,227) = .69$, $p = .408$, $\eta p^2 = .00$), on justifications ($F(1,227) = 3.07$, $p = .081$, $\eta p^2 = .01$) or on other utterances ($F(1,227) = .29$, $p = .588$, $\eta p^2 = .00$). Figures

Figure 1 shows the means of response types per interview medium.

Effect of Interview Medium on Interview Performance

To investigate the effect of the interview medium on interview performance (Research Question 2), we computed an independent-samples *t* test. Results showed no significant effect of interview medium on performance, $t(216.10) = -1.74$, $p = .083$, Cohen's $d = -.23$, 95% CI [--.50, .03], (AVIs: $M = 2.97$, $SD = .86$, FTF: $M = 2.78$, $SD = .74$).

Effect of Interview Medium on Behavioral Interview Criterion Validity

To assess the criterion validity of past-behavior questions (Research Question 3), we conducted two moderation analyses. The first analysis investigated the relationship between overall interview performance and exercise performance at the layoff with the interview condition as a moderator. The second moderation analysis investigated the relationship between the performance at the past-behavior question about the layoff with the exercise performance at the layoffs, with the interview condition as a moderator.

For the first moderation analysis investigating the relationship between overall interview performance and exercise performance, we-centered overall interview performance at the grand mean. Results revealed a significant effect of interview performance ($B = 1.28$, $t = 2.60$, $p = .010$) on exercise performance. There was no effect of interview condition ($B = 0.12$, $t = 0.50$, $p = .618$), nor a moderating effect of interview condition ($B = -0.57$, $t = -1.90$, $p = .059$), $R^2 = .050$, $F(3, 214) = 3.73$, $p = .012$. This shows that interview performance is a valid predictor of exercise performance, and interview medium does not affect this relationship.

For the second moderation analysis focusing on the specific relationship between the performance at the past-behavior question about the layoff with the exercise performance at the layoffs, we centered the performance at the past-behavior question about the layoff at the

grand mean. Results showed a significant effect of performance at the past-behavior question about the layoff (B = 0.82, t = 2.30, p = .023) on exercise performance at the layoffs. There was no effect of the interview condition (B = 0.05, t = 0.21, p = .838), nor a moderating effect of interview condition (B = -0.29, t = -1.23, p = .218), R^2 = .063, F(3, 214) = 4.76, p = .003. This indicates that the performance at the past-behavior question about the layoff is a valid predictor of exercise performance at the layoff and the interview medium does not affect this relationship.

Discussion

This research compared the effect of AVIs and FTF interviews on applicants' responses to past-behavior questions (Research Question 1), their corresponding interview performance (Research Question 2), and the criterion validity of past-behavior questions (Research Question 3). We found that participants provided more extended responses in AVIs than in FTF interviews by a factor of about 1.67. However, they do not differ in the rate of narrative answers (i.e., stories and pseudo-stories) nor in decontextualized assertions about self-descriptions and justifications. The only difference is that participants talked more about their values and opinions in AVIs than in FTF interviews. This difference could stem from the on-screen timer in the AVIs indicating the time left to respond. Some participants may have reached the end of their response before the time was up and may have felt pressured to extend their response with additional information to fill the remaining time (Orji et al., 2024). Adding information about values and opinions in the remaining time may be explained by the fact that stories are typically recounted to perform a social action like accounting for one's behavior or illustrating a characteristic of one of the protagonists (Mandelbaum, 2013). As such, narrators often make the point (i.e., the moral) of the story at the end of the telling. If audiences fail to react appropriately, they may repeat or emphasize the point (Bavelas et al., 2000). In the AVI setting,

a still-running timer may have a similar effect as an unresponsive audience in a FTF setting. This may induce participants to talk more and make a more explicit link between the story and how it illustrates their personal characteristics, thereby producing more statements about their values and opinions.

There was no significant difference in interview performance between FTF interviews and AVIs. This may partly reflect the similarity in participants' response content across media. Previous research comparing interview formats found higher performance in FTF interviews compared to videoconferencing (Melchers et al., 2021), while AVIs yielded higher performance than videoconferencing (Langer et al., 2017). The only study comparing FTF interviews with AVIs reported comparable interview performance across both media (Kleinlogel et al., 2023). Although our findings align with this research, further investigation into applicants' performance in AVIs (versus FTF interviews) is needed to reinforce and complete our understanding of the interview medium effects.

Regarding the effect of the interview medium on past-behavior question criterion validity, the overall interview performance (i.e., responses to the three past-behavior questions) predicted exercise performance as measured by the role-plays of the layoff. There was no effect of the interview medium. Moreover, the performance at the past-behavior question about the layoff predicted the exercise performance at the layoff and there was no effect of the interview medium. While behavioral interviewing is widely considered a valid predictor of future job performance in FTF interviews (e.g., Motowidlo et al., 1992; Huffcutt et al., 2004; Taylor & Small, 2010; Hartwell et al., 2019), our results suggest that its validity holds also in AVIs. This aligns with a recent study reporting good criterion validity of past-behavior questions in AVIs using another kind of performance measure, supervisor evaluation (r = .24; Liff et al., 2024).

This research carries practical implications for both organizations and applicants. First, on-screen past-behavior questions in AVIs lead to interview performance comparable to FTF

interviews. This suggests that AVIs can be a fair interview method. The comparable performance across media also responds to applicants' concerns about the lower chance of performing well in AVIs (Kleinlogel et al., 2023). Second, applicants' responses do not substantially differ across interview media, although they speak more and talk more about their values and opinions at the end of their AVI responses (Orji et al., 2024). However, this may depend on the way AVIs are configured (e.g., time allotted for response, see below). Finally, the advantages of behavioral interviewing in terms of criterion validity (Hartwell et al., 2019) seem to extend to AVIs, confirming previous research (Liff et al., 2024).

Our study has some limitations. First, it occurred in the lab, thus limiting generalizability to real, high-stakes interviews or layoffs, despite our attempts to incentivize participants. However, an experimental approach enables investigating causal relations between variables of interest (Bless & Burger, 2016). Also, it would have been challenging to control the past job-related situation (what people did) and ask about it in a job interview (what people say they did) in real-life interviews, especially when comparing differences in interview medium while holding other factors constant, which is a key advantage of experiments (Falk & Heckman, 2009). We thus advocate for more experimental research, given the current dearth of knowledge on differences in applicant responses to past-behavior questions between AVIs and FTF interviews.

A second limitation concerns the AVI configuration. While multiple parameters can be manipulated (Lukacik et al., 2022; Roulin et al., 2023), we chose a fixed set of those parameters to make our AVIs comparable to FTF interviews. For example, AVI participants had a 20-second time window to read each interview question. This was intended to standardize the reading process, but potentially allowed fast readers to start preparing a response or to initiate their response before the timer ended. The lack of control about who initiated their response first and who waited until the countdown ended introduces a potential source of variability

between the FTF and AVI conditions. This may have added noise in our comparison with FTF interviews, where participants listen to the recruiter's questions, sometimes asked for clarifications, and then proceeded to answering the question. At the same time, participants in the FTF condition may have already started preparing their response while the recruiter was still asking the question, as is typical in conversation (Levinson, 2016). Despite this limitation, our results align with previous research on applicants' interview performance when having no preparation time (Kleinlogel et al., 2023).

Future studies should investigate the generalizability of these findings across different job positions and domains. Moreover, they should assess the impact of different AVI parameters on applicants' responses to past-behavior questions and performance, such as examining the effect of the time allotted to respond, and determining whether such parameters have boundary conditions for optimal performance. Also, one promising avenue may involve parametrizing AVIs to facilitate participants' recall of episodes of maximal rather than typical performance (Huffcutt et al., 2024), e.g., by priming participants with question formulation and fine-tuning preparation time.

Conclusion

AVIs featuring past-behavior questions represent a promising alternative to FTF interviews but raise questions about their impact on applicants' responses, interview performance, and criterion validity. Our findings revealed that the interview medium has little effect on these outcomes. Because very little research exists on these questions, further research should continue to explore the effects of different AVI designs on applicants' responses, interview performance and on AVI criterion validity.

Acknowledgement

This work was supported by the Swiss National Science Foundation (SNSF) under Grant 10521C_197479.

Declaration statement

The authors report no conflict of interest.

Data availability statement

The data supporting this study findings will be openly available on OSF repository at the time of publication.

References

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. *Journal of Business and Psychology*, *29*(4), 593-604. https://doi.org/10.1007/s10869-014-9350-0

Basch, J. M., Melchers, K. G., Kegelmann, J., & Lieb, L. (2020). Smile for the camera! The role of social presence and impression management in perceptions of technology-mediated interviews. *Journal of Managerial Psychology 35*(4), 285-299. https://doi.org/10.1108/JMP-09-2018-0398

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology 79*(6), 941-952.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). {e}ffectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Bless, H., & Burger, A. M. (2016). A Closer Look at Social Psychologists' Silver Bullet: Inevitable and Evitable Side Effects of the Experimental Approach. *Perspectives on Psychological Science*, *11*(2), 296-308. https://doi.org/10.1177/1745691615621278

Bourdage, J. S., Roulin, N., & Tarraf, R. (2018). "I (might be) just that good": Honest and deceptive impression management in employment interviews. *Personnel Psychology*, *71*(4), 597-632. https://doi.org/10.1111/peps.12285

Breuer, K., Nieken, P., & Sliwka, D. (2013). Social ties and subjective performance evaluations: an empirical investigation. *Review of managerial Science*, *7*, 141-157. https://doi.org/10.1007/s11846-011-0076-3

Brosy, J., Bangerter, A., & Mayor, E. (2016). Disfluent responses to job interview questions and what they entail. *Discourse Processes*, *53*(5-6), 371-391. https://doi.org/10.1080/0163853X.2016.1150769

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. *European Journal of Work and Organizational Psychology*, *29*(3), 330-343. https://doi.org/10.1080/1359432X.2019.1704265

Burroughs, W. A., & White, L. L. (1996). Predicting sales performance. *Journal of Business and Psychology*, *11*(1), 73-84. https://doi.org/10.1007/BF02278257

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, *50*(3), 655-702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Ciphr. (2023). *Are video interviews the future of recruitment?* https://www.linkedin.com/pulse/video-interviews-future-recruitment-ciphr/

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology*, *13*(4), 485-506. https://doi.org/10.1023/A:1022914803347

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *science*, *326*(5952), 535-538. https://doi.org/10.1126/science.1168244

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (3 ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Hartwell, C. J., Johnson, C. D., & Posthuma, R. A. (2019). Are we asking the right questions? Predictive validity comparison of four structured interview question types. *Journal of Business Research*, *100*, 122-129. https://doi.org/10.1016/j.jbusres.2019.03.026

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

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Heimann, A. L., Ingold, P. V., & Kleinmann, M. (2020). Tell us about your leadership style: A structured interview approach for assessing leadership behavior constructs. *The Leadership Quarterly*, *31*(4), 101364.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment*, *12*(3), 262-273. https://doi.org/10.1111/j.0965-075X.2004.280_1.x

Huffcutt, A. I., Howes, S. S., Murphy, D. D., & Murphy, S. A. (2024). Enhancing consistency of maximal responding in behavior description interviews: An exploration of priming and response length. *Personnel Assessment and Decisions*, *10*(1), 1. https://doi.org/10.25035/pad.2024.01.001

Huffcutt, A. I., & Murphy, S. A. (2023). Structured interviews: moving beyond mean validity…. *Industrial and Organizational Psychology*, *16*(3), 344-348. https://doi.org/10.1017/iop.2023.42

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158-168). Sage Publications.

Kessler, R. (2006). *Competency-based interviews: Master the tough new interview style and give them the answers that will win you the job*. Frankilin Lakes : Career Press.

Kleinlogel, E. P., Schmid Mast, M., Jayagopi, D. B., Shubham, K., & Butera, A. (2023). "The interviewer is a machine!" Investigating the effects of conventional and technology-mediated interview methods on interviewee reactions and behavior. *International Journal of Selection and Assessment*, *31*, 403-419. https://doi.org/10.1111/ijsa.12433

Komsta, L., & Novomestky, F. (2022). *moments: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. Retrieved from https://CRAN.R-project.org/package=moments

Krajewski, H. T., Goffin, R. D., McCarthy, J. M., Rothstein, M. G., & Johnston, N. (2006). Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future? *Journal of Occupational and Organizational Psychology*, *79*(3), 411-432. https://doi.org/10.1348/096317905X68790

Langer, M., König, C. J., & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. *International Journal of Selection and Assessment*, *25*(4), 371-382. https://doi.org/10.1111/ijsa.12191

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research*, *39*(2), 329-358. https://doi.org/10.1207/s15327906mbr3902_8

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational and Organizational Psychology*, *73*(1), 67-85. https://doi.org/10.1348/096317900166886

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, *20*(1), 6-14. https://doi.org/10.1016/j.tics.2015.10.010

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001173

LinkedIn. (2019). *The Tactic This Expert Uses to Assess Soft Skills* Retrieved 29.08.2023 from https://www.linkedin.com/business/talent/blog/talent-acquisition/tactic-this-expert-uses-to-assess-soft-skills

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), 1-15. https://doi.org/10.1016/j.hrmr.2020.100789

Mandelbaum, J. (2013). Storytelling in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 492-507). Chichester, UK: John Wiley & Sons.

Mejia, C., & Torres, E. N. (2018). Implementation and normalization process of asynchronous video interviewing practices in the hospitality industry. *International*

Journal of Contemporary Hospitality Management, *30*(2), 685-701. https://doi.org/10.1108/IJCHM-07-2016-0402

Melchers, K. G., Petrig, A., Basch, J. M., & Sauer, J. (2021). A comparison of conventional and technology-mediated selection interviews with regard to interviewees' performance, perceptions, strain, and anxiety. *Frontiers in Psychology*, *11:603632*. https://doi.org/10.3389/fpsyg.2020.603632

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., & Vaughan, M. J. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology*, *77*(5), 571-587. https://doi.org/10.1037/0021-9010.77.5.571

Orji, K., Bangerter., A., Germanier, E., Renier, L. A., Schmid Mast, M., H., M. , & Garner, P. N. (2024). *Extended Responses in Asynchronous Video Interviews: Investigating Frequency, Content, and Interview Outcomes [Manuscript in preparation]*. University of Neuchâtel.

R Core Team. (2021). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Ralston, S. M., Kirkwood, W. G., & Burant, P. A. (2003). Helping interviewees tell their stories. *Business Communication Quarterly*, *66*(3), 8-22. https://doi.org/10.1177/108056990306600303

Richter, M., König, C. J., Koppermann, C., & Schilling, M. (2016). Displaying fairness while delivering bad news: Testing the effectiveness of organizational bad news training in the layoff context. *Journal of Applied Psychology*, *101*(6), 779–792. https://doi.org/10.1037/apl0000087

Roulin, N., Bangerter, A., & Wüthrich, U. (2012). *Réussir l'entretien d'embauche comportemental: La méthode pour identifier et sélectionner les futurs employés performants*. Bruxelles: De Boeck Professionals.

Roulin, N., Wong, O., Langer, M., & Bourdage, J. S. (2023). Is more always better? How preparation time and re-recording opportunities impact fairness, anxiety, impression management, and performance in asynchronous video interviews. *European Journal of Work and Organizational Psychology*, *32*(3), 333-345. https://doi.org/10.1080/1359432X.2022.2156862

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

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology*, *77*(5), 629. https://doi.org/10.1037/0021-9010.77.5.629

Schneider, L., Powell, D. M., & Bonaccio, S. (2019). Does interview anxiety predict job performance and does it influence the predictive validity of interviews? *International Journal of Selection and Assessment*, *27*(4), 328-336. https://doi.org/10.1111/ijsa.12263

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*(2), 149-155. https://doi.org/10.1037/h0047060

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, *80*(5), 587-606. https://doi.org/10.1037/0021-9010.80.5.587

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology*, *75*(3), 277-294. https://doi.org/10.1348/096317902320369712

Turner, T. S. (2004). *Behavioral interview guide: A practical, structured approach for conducting effective selection interviews*. Victoria, B.C. : Trafford Publishing.

Van Iddekinge, C. H., Raymark, P. H., Eidson, J., Carl E, & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance*, *17*(1), 71-93. https://doi.org/10.1207/S15327043HUP1701_4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the {tidyverse}. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Tables

Table 1.

Descriptive Statistics of Study Variables by Interview Condition.

	FTF			AVI		
	n	M	SD	*n*	M	SD
Exercise performance	109	18.71	1.84	110	18.92	1.69
Layoff Question Performance	112	2.98	1.08	111	3.33	0.97
Interview Performance	112	2.78	0.74	111	2.97	0.86
Utterance Counts	114	62.42	30.77	115	105.71	45.09
Stories	114	34.35	24.06	115	51.90	33.48
Pseudo-stories	114	13.01	13.15	115	22.08	18.89
Values/Opinion	114	10.62	10.68	115	24.59	20.04
Self-description	114	0.92	1.95	115	1.70	2.39
Justification	114	1.43	1.99	115	1.83	2.13
Other Utterances	114	1.99	2.16	115	3.76	3.32

Note. FTF = face-to-face interviews, AVI = asynchronous video interviews.

Table 2

Mean, Standard Deviation and Correlations Between Main Study Variables.

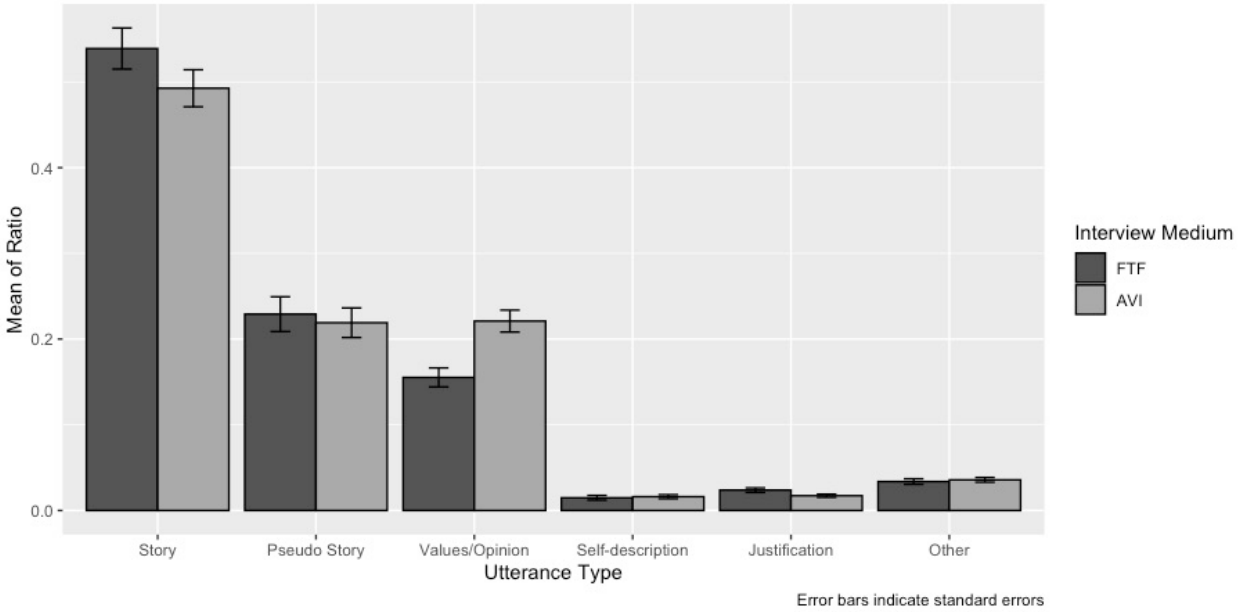
	N	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1 Gender	229	1.39	0.49												
2 Age	229	21.55	3.09	.05											
3 Job Interview Experience	229	1.92	2.78	.06	.58 ***										
4 Interview Condition	229	1.50	0.50	.00	.00	.02									
5 Exercise performance	219	18.82	1.76	-.19 ***	.04	.06	.06								
6 Stories	229	51.53	24.43	.14 *	-.04	-.04	.27 ***	.05							
7 Pseudo-stories	229	22.4	20.07	-.17 **	-.02	-.07	.25 ***	.11	-.82 ***						
8 Values/Opinion	229	18.88	13.25	.02	.05	.09	.37 ***	.08	-.47 ***	-.08 ***					
9 Self-description	229	1.55	2.75	-.02	-.07	-.07	.15 *	.03	-.23 ***	.10 **	.06				
10 Justification	229	2.04	2.47	.01	-.05	-.07	.08	.02	.00	-.05 ***	.05	.06			
11 Other Utterances	229	3.48	3.20	.01	.05	.08	.28 ***	.08	-.21 **	.05 ***	-.16	.05	-.12		
12 Layoff Question Performance	223	3.16	1.04	-.08	-.01	.02	.19 **	.22 ***	.21 **	-.10	-.12	-.16 *	-.13	-.13	
13 Interview Performance	223	2.87	0.81	-.01	.00	.01	.10	.19 **	.01	-.02 ***	.09	-.02	-.10	-.16 *	.66 **

Note. Pearson correlation with pairwise deletion; Gender: 1 = female, 2 = male; Interview Condition: 1 = FTF, 2 = AVI; * $p < .05$; ** $p < .01$; *** $p < .001$

Figures

Figure 1

Effect of Interview Medium and Response Type on the Utterance Ratio.



Note. $N = 229$ ($N_{FTF} = 114$, $N_{AVI} = 115$); FTF = face-to-face interviews, AVI = asynchronous video interviews; error bars indicate ± 1 SE.

IV Appendix 4: Study 4

Germanier, E.*, Mutian, H.*, Rufai, A. M., Garner, P. N., Bangerter, A.,Renier, L. A., Schmid Mast, M., & Orji, K. (2024). *Identifying Storytelling in Job Interviews Using Deep Learning* [Manuscript in preparation]. Institute of Work and Organizational Psychology, University of Neuchâtel

* shared first authorship

Status: The manuscript is submitted to a journal. The version in this thesis is the manuscript initially submitted.

Identifying Storytelling in Job Interviews Using Deep Learning

Elisabeth Germanier*[1], Mutian He*[2,3], Amina Mardiyyah Rufai[2],
Philip N. Garner[2], Adrian Bangerter[1], Laetitia A. Renier[4],
Marianne Schmid Mast[4], and Koralie Orji[1]
[1]Institute of Work and Organizational Psychology, University of Neuchâtel, Switzerland
[2]Idiap Research Institute, Martigny, Switzerland
[3]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[4]Department of Organizational Behavior, University of Lausanne, Switzerland

Structured interviews often include past-behavior questions inviting applicants to recount a past work experience. While optimal responses to these questions should take the form of a story, applicants struggle to produce them extemporaneously. Asynchronous video interviews (AVIs) present new opportunities for job interview coaching, which can incorporate artificial intelligence to analyze video-recorded responses and deliver personalized feedback. Following this demand, we explore the potential of audio-based deep-learning models to identify storytelling and other, sub-optimal responses (pseudo-stories, decontextualized assertions) from interview recordings. Using data from 254 mock interviews featuring three past-behavior questions, we developed models to determine the utterance type, considering different scenarios and labeling schemes of varying granularity. We further applied multiple techniques to improve the model accuracy. Findings show that our models achieve satisfactory performance when enhanced with audio information and enriched with longer context (best accuracy: 77.67%). In the end, we discuss the results, implications, and future research directions.

Keywords: Deep learning, storytelling, past-behavior questions, selection interviews

Introduction

Behavioral job interviews are a best practice to assess applicants' competencies (Janz, 1989). They often use past-behavior questions (*Please tell me about an occasion where you had to deal with an unsatisfied client)*, which invite applicants to give a detailed account of a past work-related event where they took action to resolve a problematic situation. Those questions thus invite applicants to tell a story (Ralston et al., 2003). Telling stories in job interviews may allow applicants to convey a good impression in recruiters' eyes (Stevens & Kristof, 1995) and increase their chances of being hired (Bangerter et al., 2014). However, applicants often struggle to produce stories on demand and may opt for sub-optimal responses (Bangerter et al., 2014; Brosy et al., 2016). Different strategies to help applicants improve their storytelling have been investigated, including coaching (Ralston et al., 2003), image-based training (Lin-Stephens et al., 2022), and providing information about the interview format or using probe questions during interviews (Brosy et al., 2020).

In recent years, new technology-driven interview formats have emerged, like asynchronous video interviews (AVIs), where applicants log into a platform from their computer or smartphone and audio- or video-record their responses to questions asked in written form or by an avatar (Lukacik et al., 2022). Due of the lack of interactivity, applicants' responses may differ from face-to-face (FTF) settings (Germanier et al., 2024). Moreover, some traditional solutions to improve applicants' responses (e.g., probe questions) become impossible in AVIs. Therefore, helping applicants respond better to past-behavior questions may be complicated by the AVI setting.

Recently, recordings generated via AVIs may also enable novel solutions for training applicants. In addition to assisting recruiters in hiring decisions (e.g., Naim et al., 2015), automatic analyses of applicants' responses via machine learning models can also support the delivery of automated feedback in training applications to help applicants improve their responses. To provide applicants with personalized feedback, verbal and nonverbal behaviors extracted from applicants' interview video recordings have been recently investigated (Gebhard et al., 2018; Hoque et al., 2013; Langer et al., 2016). Particularly, Bangerter et al. (2023) used machine learning approaches to automatically identify storytelling responses to past-behavior questions from transcripts of FTF interviews. However, their approach is constrained by

* Equally contributed and share the first authorship.

the need for manual transcription, which prevents the implementation of these models in training platforms. Plus, more recently, powerful end-to-end machine learning using deep neural networks, known as deep learning (DL), has demonstrated strong performance in making inference from text, audio and visual information. Systems based on DL have been implemented in a broad range of applications including chatbots (OpenAI, 2023), speech recognition (Radford et al., 2023), spoken language understanding (Qin et al., 2021), and interview analysis (Koutsoumpis et al., 2024; Rahman et al., 2021).

Building on these aforementioned works and the recent progresses in DL , this study aims to developp DL models cabable of automatically identifying storytelling responses as well as other sub-optimal types of response directly from audio-recorded responses. Our study goes beyond previous work in (1) focusing on the content of participants' responses, (2) evaluating performance of DL models under the pretraining+fine-tuning paradigm rather than traditional machine learning (e.g., random forest), and (3) working directly from audio recordings, which brings us a step closer to practical implementation in real application scenarios. To do so, we apply DL models to identify response type from audio-recorded responses of participants in mock job interviews in an experimental setting with FTF interviews and AVIs. The goal of this research is to pave the way for developing models to implement in AI-powered training platform that provides automatic feedback on applicants' responses to past-behavior questions from the audio recordings.

Applicants' Responses to Past-Behavior Questions

Job interviewing is a common procedure for recruiters to assess applicant characteristics such as personality and competencies (Huffcutt et al., 2001). Behavioral interviews are a best practice in interviewing. They often feature past-behavior questions, which are valid predictors of future job performance. Past-behavior questions invite participants to describe their behavior in past work-related situations (Campion et al., 1997; Janz, 1989). Applicants' responses to such questions should take the form of narratives (i.e., storytelling) (Ralston et al., 2003). Indeed, storytelling is a good way to create a positive impression, which provides an engaging and potentially more memorable response (Bangerter et al., 2014; Brosy et al., 2016; Stevens & Kristof, 1995), thus increases hirability ratings (Bangerter et al., 2014). Practical interview guides often advise applicants to use the STAR mnemonic to organize their stories: Start with a brief description of the situation (S), explain the task and actions undertaken (TA), and describe results obtained (R) (Kessler, 2006). For example, regarding the past behavior question:

> Can you tell me about a situation where one of your employees didn't complete a task on time? How did you react to this problem?

a storytelling (S) response from our corpus (translated from French) is:

> *then uh yes it did happen uh during a project to be handed in uh for uh the courses* (SS)*/ and um what happened is that I uh discussed with uh the various uh collaborators who hadn't handed in their work on time* (STA)*/ and so we sent an e-mail uh* [laughs] *to the professor* (STA)*/ to extend the deadline* (STA)*/ to find a way* (STA)*/ basically we used uh I tried to communicate the best I could with them* (STA)*/ and tell them what was wrong* (STA)*/ and what solutions we could perhaps uh put in place* [laughs] (STA)

This example constitutes a story because it describes a unique episode (*it did happen uh during a project*) in the past, one marker of this being the use of the past tense. In this example, the response has been segmented into utterances and each utterance has been labelled as being about the Situation (one utterance), Task and Actions (seven utterances), or Results (zero utterances). Besides the words constituting each utterance, the response transcript also shows evidence of filled pauses ("uh") and paraverbal behavior ("laughs").

Applicants do not always spontaneously provide storytelling responses when asked past-behavior questions, and often provide sub-optimal responses such as general descriptions of generic situations (pseudo-stories, P), as in the following response to the same question:

> *hmm for example a colleague* (PS)*/ well I try if I'm able to save the situation in quotes* (PTA)*/ well I give him my help* (PTA)*/ and so that everyone gets out of it* (PTA)*/ and but if I can't do anything* (PTA)*/ well I couldn't do anything* (PTA)*/ well then I wouldn't be the one to uh try to sink the person or whatever* (PTA)*/ I'll always try to help him* (PTA)

This example constitutes a pseudo-story because it describes a generic situation using the present tense without any individualizing information (e.g., specific time and/or location). Another type of sub-optimal response consists of decontextualized assertions about oneself, including values/opinions (VO), self-descriptions (SD), or justifications (JS), as in the following example from our corpus. Given the question:

> Can you tell me about a situation where you were responsible for carrying out several tasks in a short space of time? Tell me how you organized yourself to deal with this situation.

the response is:

> *okay uh well the* [sighs] *multitasking uh of course she uh doing several tasks uh at the same time is not my forte* (SD)/ *but uh I think we w- we can all do several tasks uh at the same time* (VO)/ *so uh me that's what- the style I prefer uh to do one at a time uh* (SD)/ *but uh um there are times when we're obliged to do uh several uh things uh at the same time* (VO)/ u*h and uh the switch and manage the tasks at the same time it's a uh it's a bit tiring for the brain* (VO)/ *but uh uh I can do it* (SD)/ *so uh it's not it's what I prefer* (SD)/ *uh but uh I can do it* (SD)

In this example, the participant does not answer the question by producing a story. Rather, they produce a series of self-descriptions and values/opinions. In one study, when asked past-behavior questions, these decontextualized responses were more likely to occur the longer applicants paused before responding (Brosy et al., 2016). Participants may have trouble finding a suitable past episode to recount (evidenced by longer pauses) and thus resort to self-descriptions and statements of values and opinions in order to start speaking, even if the response is sub-optimal.

Table 1

Abbreviations by Utterance Type

Utterance Type	Abbreviation
Story	S
Situation	SS
Task-Action	STA
Result	SR
Pseudo-Story	P
Situation	PS
Task-Action	PTA
Result	PR
Values/Opinion	VO
Self-description	SD
Justification	JS
Other	OR

These examples illustrate the difficulties in producing stories in response to past-behavior questions and may constitute an important factor affecting the validity of the interview procedure (Huffcutt & Murphy, 2023). Applicants' responses can be improved with job interview training and recruiters' help. A first approach involves an instructor (e.g., coach or career advisor) critiquing the response, helping the applicant construct a corrected version of the narrative that is then discussed within a training group (Ralston et al., 2003). A second approach consists of extended training with image-based intervention wherein interviewees create images associated with their examples (Lin-Stephens et al., 2022). However, these procedures are time-consuming and can be difficult for applicants to implement (e.g., the lack of access to a coach). Alternatively, recruiters may provide information about the type of question before the interview or use probe questions during the interview to help applicants improve the content of their responses (Brosy et al., 2020). While probe questions are effective in increasing storytelling responses and balancing the narrative elements of stories (STAR), providing information prior to the interview situation is not (Brosy et al., 2020).

The Rise of Asynchronous Video Interviews

In-person job interviews are being increasingly supplemented by asynchronous job interviews (AVIs). AVIs are more flexible, so that recruiters and applicants do not need to meet in space and time anymore (Basch et al., 2020). Applicants connect to a platform and video-record their responses to questions which are displayed on-screen or asked by an avatar. Recruiters, or, more recently, algorithms, evaluate applicants' responses based on the video-recording (Lukacik et al., 2022).

Since AVIs constitute different interactional situations, applicants' responses to past behavior questions are likely to differ. Compared to FTF interviews, applicants' storytelling responses feature similar proportions of STAR narrative elements in AVIs (Germanier et al., 2023). However, applicants talk more about their values and opinions in AVI settings (Germanier et al., 2024). Such expression of personal values may not constitute relevant responses to past-behavior questions, resulting in lower chances of being hired (Orji et al., 2024). To help with this, video-based training has been developed, by explaining for instance what constitutes good behaviors leading to higher performance, including storytelling (Roulin, Pham, & Bourdage, 2023). An alternative training approach for effective storytelling could involve AI-powered systems within coaching platforms or software to automatically provide tailored feedback to applicants based on recordings of their responses.

Analyzing Interview Responses of Applicants with Machine Learning and Deep Learning

Along with the advent of AVIs, significant progress has been made in machine learning (ML). Recent ML techniques can make powerful predictions of labels (e.g., human-coded labels like storytelling or other types of utterances in the present case) from observations (e.g., human speech). ML can thus assist human decision-makers (Liem et al., 2018). To generate robust and accurate predictions, ML models typically follow a two-step process. First, they are trained to

adapt to the data. This first step can be typically accomplished through supervised learning, where the model is provided with human-coded labels, which serve as the *ground truth* that the model uses to learn (Liem et al., 2018). Second, the trained model is tested and evaluated on previously unseen data to see how well it works with any newly presented data it will encounter when applied to practical use. Of note, the quality of the ML prediction highly depend on the quality of the training data (Geiger et al., 2020)

Machine learning has been applied in AVIs to predict applicants' personality traits or hiring recommendations based on automatically extracted verbal, paraverbal and visual features as well as human annotations (e.g., Chen et al., 2017; Hickman et al., 2022; Holtrop et al., 2022; Naim et al., 2015; Nguyen et al., 2014; Rupasinghe et al., 2016). In the context of job interview training, ML applications can provide applicants with automatic feedback to improve their interview performance. One study developed an application where interviewees interact with a conversation avatar coach to get feedback (Hoque et al., 2013). Another study implemented automatic feedback on interviewees' behavior into serious games for job interview training (Gebhard et al., 2018). Langer et al. (2016) also investigated how real-time feedback on nonverbal behaviors in avatar training job interviews improves interviewee performance.

The above systems deliver feedback on nonverbal or paraverbal aspects of applicants' responses. One system delivers feedback on "weak language" (i.e., filler words) (Hoque et al., 2013). Currently, no systems we are aware of are capable of delivering feedback tailored to the substantive content of participants' responses. This is an important gap because content is one of the most important factors affecting interview performance, being more important than nonverbal or paraverbal behavior (Rasmussen, 1984; Riggio & Throckmorton, 1988). Further, content is particularly important in evaluating or improving responses to past-behavior questions. In a first step in this direction, Bangerter et al. (2023) used ML to identify storytelling in applicants' responses in FTF interviews. Based on interview transcripts, they automatically extracted semantic features (i.e., LIWC; Pennebaker et al., 2015), word counts, and TF-IDF (a measure of a word's importance in a text). Using various ML algorithms (e.g., random forest), they predicted whether a response consisted of a story or not considering the amount of different STAR narrative elements in responses. Their findings suggested that it is feasible to identify a particular utterance type, storytelling, in applicants' responses using ML. A limitation of that study is the reliance on transcribed speech, which restricts the applicability of the approach to a real training platform setting. Indeed it is not feasible to manually transcribe the audio recordings prior to the ML analyses in a scalable application.

More sophisticated deep neural techniques may enable progress. These techniques have also been leveraged in AVI settings, e.g., to predict personality using convolutional neural network (Suen et al., 2019), or hirability using hierarchical gated recurrent unit (Hemamou, Felhi, Vandenbussche, et al., 2019), though still using handcrafted features, with attention weights used to identify critical timespans (Hemamou, Felhi, Martin, & Clavel, 2019). While starting from the landmark work of BERT (Devlin et al., 2019), the state-of-the-art in DL is characterized by the *pretraining+fine-tuning paradigm*. Under this paradigm, a large language model is first trained upon a very large dataset of unlabeled text or speech, using *self-supervised tasks* such as doing cloze questions. Since such tasks do not need human coded labels, a large amount of data from books and the Internet can be used at little cost. Furthermore, doing well on such language tasks requires the model to learn general knowledge and capability to understand language and produce feature representations that encapsulate the structure and semantics of the input texts. In simpler terms, the model must convert words into mathematical representations (vectors) that reflect both how the text is structured (e.g., grammar) and what it means (e.g., context and word associations). By fine-tuning various publicly available pretrained models on local machines, such language knowledge and capability can then be transferred to more specific downstream or target tasks. Examples are sentiment classification, question-answering (Devlin et al., 2019), spoken language understanding (Qin et al., 2021), and so should be interview analysis. Given knowledge learned from a large amount of pretraining data, the demand of the data size on the target task can be greatly reduced. This approach is essential in high-performance DL systems nowadays, especially when the high cost of collecting realistic data (as in the case of interview responses) becomes a restriction of the target dataset. Attempts have been made to incorporate pretrained models into interview analysis, such as using them to process interview transcripts for predicting hirability (Rahman et al., 2021). However, there is still much to explore about the potential of end-to-end DL with pretrained models in different modalities (including both language and speech) in analyses of applicants' response content.

This Study

Applicants' responses to past-behavior questions are predictive of their future job performance. However, applicants are not always able to produce appropriate responses (i.e., stories) to these questions in real time. This may interfere with correct assessments of their personal characteristics and explain in part the high variability in validity of past-behavior questions (Huffcutt & Murphy, 2023). Procedures for coaching applicants show potential but are often expensive or time-consuming. AVIs may further complicate responding to past-behavior questions. However, the automatically generated

video recordings in AVIs hold potential for automatic analysis of participants' response behaviors and thus the provision of automatically generated feedback tailored to participants' specific responses, thus bypassing the need for coaching or at least potentially complementing coaching interventions (Bangerter et al., 2023).

Accordingly, this study used DL models with the goal of predicting storytelling components (S/TA/R) and other utterance types such as pseudo-story components (S/TA/R), values/opinions, self-descriptions, and justifications to past-behavior interview questions from audio recordings of participants answering past-behavior questions in mock job interviews. Participants' responses were labeled by human coders. These labels were used as the ground truth for training models.

This study goes beyond previous work (e.g., Bangerter et al., 2023) in several respects. First, we used the state-of-the-art DL technique under the pretraining+fine-tuning paradigm, which is essential in the current situation where the training set of participants' responses is relatively scarce and linguistically complex. Specifically, we used the multilingual version of the recent pretrained textual model RoBERTa (Liu et al., 2019) and speech model Wav2Vec2 (Baevski et al., 2020), in combination with the state-of-the-art pretrained speech recognizer Whisper (Radford et al., 2023). Second, we predicted a large set of labels (i.e., 10) that constitute a comprehensive and fine-grained array of applicants' potential responses to past-behavior questions. Third, we considered a range of different modeling techniques to identify the impact of various factors so as to cater to practical needs with optimal performance.

Using these pretrained DL models, our modeling followed a three-stage development. In the first stage, we conducted *baseline assessments*. We compared the performance of the models under the evaluation scenarios using human transcripts (i.e., *transcription-based* scenarios), or using automatic speech recognition (ASR) models that generate transcripts to mirror practical cases when only the audio is available(i.e., *audio-based* scenarios). In addition to the original labeling scheme with 10 labels, we also assessed two higher-level labeling schemes with fewer categories. This is because prediction of labels that were less frequent 1 or that were difficult to discriminate, especially with limited context, may be challenging and less accurate. More importantly, our objective is to detect storytelling, and it could be useful to consider simplified schemes that are more focused on this goal.

The second stage consisted of *information enrichment* in training of audio-based models. We considered the effectiveness of the model trained upon the human transcripts only, versus models trained with additional information from the audio. Such techniques include using transcripts produced by ASR during training, as well as directly introducing the audio as part of the model input. We explored these scenarios because human transcripts, although more accurate and generally used for model training, are not available in real cases based on interview recordings where we can only access ASR transcripts. Furthermore, paraverbal cues (e.g., intonation, prosody) might be also informative for this semantic-centric classification, but are not well represented in transcripts.

The third stage consisted of *context enlargement*. Given the significant variation in utterance length and the prevalence of relatively short utterances, it is often challenging to accurately determine the utterance type. For instance, an utterance describing an action may be classified as STA if the exact time of the event is mentioned in the previous utterances but as PTA otherwise. Therefore, incorporating more context of the target utterance into the models can enhance their performance. Hence we created two different techniques to incorporate more context, namely coalescence and expansion. Coalescence involves merging adjacent utterances of the same type, while expansion directly enlarges the context visible to the model.

Method

Participants

There were 254 French-speaking undergraduates (59.8% women, M_{age} = 21.60, SD = 3.07). Fifty-two % of participants held a high school diploma, 48.8% had a university degree (bachelor: 39% and master: 9.8%) and 1.3% held other types of diplomas. Participants had already experienced 1.94 job interviews on average (SD = 2.79). Participants were offered monetary compensation consisting of a fixed sum and a performance-based bonus.

Procedure

Participants engaged in a two-stage lab experiment. In the first stage, they started completing a consent form and a personality questionnaire (data not reported here). Then, they performed a work sample (i.e., two role plays) adapted from Richter et al. (2016). After each role-play, participants self-rated their performance (data not reported here). About one week later, participants returned for a mock job interview for a managerial job position, randomly assigned to either FTF or AVI conditions. Once the job interview was completed, participants filled out a final short questionnaire (data not reported here). Throughout both the work sample and the job interview, participants were audio and video recorded using front and side cameras, along with microphones.

All participants were asked the same four questions. The first question was about self-presentation asking "Could you please introduce yourself and summarize your background in a few words?" The second and third questions were past-behavior questions targeting managerial competencies (e.g., multitasking: "Can you describe a situation where you were

responsible for completing several tasks within a short period? How did you organize yourself to manage this situation?"). The fourth question was also a past-behavior question that targeted the role play of the previous week (i.e., "Can you tell us about a recent situation in which you had to give bad news to someone?"). In the FTF condition, participants were interviewed by a mock recruiter trained to follow a script. After the experimenter briefed them about the job interview and left the room, the recruiter greeted the participant, started with a brief overview of the interview, and continued asking the questions before concluding the interview. In the AVI condition, participants used the same AVI platform as in Roulin, Wong, et al. (2023)'s study to read and respond to on-screen questions. Following the experimenter's instructions, participants had up to 20 seconds to read each question and up to 5 minutes to answer. They had the flexibility to start and finish their responses before the respective timers ended, applying to all interview questions.

Data Preparation

Transcription and Preprocessing

All participants' responses to the past-behavior questions were transcribed manually, including disfluencies (*uh*, *um, hmm*), hesitations, repetitions, and word truncation. Their responses were then segmented into shorter units, here utterances corresponding to a subject, verb, and object structure (Bangerter et al., 2014).

Regarding the utterance lengths, there was a small but significant difference between FTF interviews and AVIs (D = 0.022, p = .016, w.r.t. word counts): Each utterance contained on average M=11.68 words (SD=5.67) or M=4.06 sec (SD=2.59) of speech for FTF interviews, and M=11.96 words (SD=6.02) or M=4.52 sec (SD=2.69) for AVIs. Thus, utterances were often quite short (see Figure 2). Also, there was a significant difference between the two interview conditions for the total length of an interview (D = 0.235, p = .002, w.r.t. word counts): For each FTF interview, there were M=1250.25 words (SD=423.53) or M=7.27 min (SD=2.31) min of non-silent audio on average, while for AVIs the number was M=1511.26 words (SD=644.69) or M=9.69 min (SD=3.82).

Manual Coding: Utterance Type

Each utterance contained in participants' responses to past-behavior questions was coded using the categories proposed by Bangerter et al. (2014), which include: story (S) considered as a "set of events related to a unique past episode, characterized by a unity of time or action"; pseudo-story (P) defined as "a description of a generic situation or recurrent set of similar situations, without unity of time or action", and decontextualized assertions about interviewees' opinions (value/opinion, VO), justification for their actions (justification, JS) or self-descriptions (self-description, SD) (Bangerter et al., 2014, p. 6). Any utterance not falling into these specific categories was coded as "others," (OR). Categories were mutually exclusive. Narrative utterances (i.e., story and pseudo-story) were further coded according to the STAR model (Kessler, 2006): Utterances providing information about the context were coded for *situational* elements (SS for stories, PS for pseudo-stories), utterances providing information about roles and actions of protagonists were classified as *task*- and *action*-related elements (STA/PTA), and those referring to the results obtained were classified as *results*-related elements (SR/PR). Two trained experts in personnel selection were involved in this coding task. Reliability was established by double-coding 30 transcripts (Fleiss' Kappa = .99, $p < .01$).

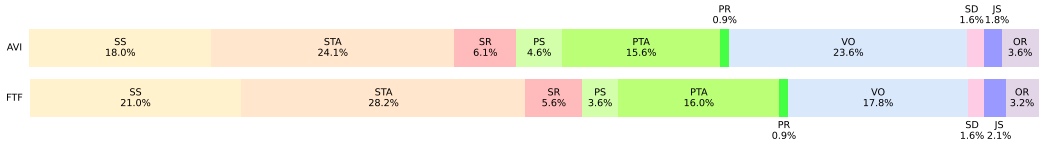
Alignment

Interview audio files collected in this study are first transcribed, and the type of each utterance is then coded from the transcript. To directly analyze the interview speech, it is necessary to associate the coded utterance type with the audiorecording of the corresponding segment, so that we can train DL models to predict the label given audio only. To achieve that, we use the standard technique known as *forced alignment* in speech processing, which aligns the transcript with the audio by identifying the exact start and end timestamps within the audio for each word in the transcript. Then it is straightforward to identify the audio segment or time interval in the recording corresponding to each utterance, and the label can be assigned to each segment accordingly.

Typically, forced alignment is carried out using acoustic models, which are also a key component of ASR systems; multiple out-of-the-box toolkits are available. However, in our data the recordings are long-form French dialogues. There are many filler words, variations in the recording quality, long silences, and sometimes overlapping speech between speakers; also the target utterances are often rather short. Long silences due to, e.g., participants thinking about their response, are particularly disturbing as they interfere with acoustic modeling. But identifying them, known as voice activity detection (VAD), is not trivial, and simple energy-based VAD (e.g., identifying a time interval as silence when the amplitude is small enough compared to other parts of the audio) fails due to variations in the dialogue and recording environment, as well as bursts of environmental noise (e.g., knocking at the table). We tried several well-known out-of-the-box toolkits for forced alignment, including Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) and WhisperX (Bain et al., 2023), but both of them fail to address these unique challenges and reach satisfying results. Therefore, we had to build a specialized forced alignment pipeline using a combination of multiple pretrained speech models and heuristics.

Figure 1

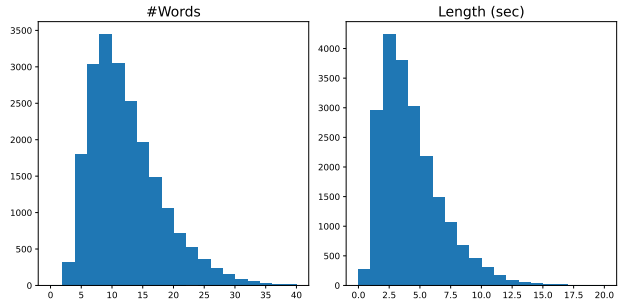
Percentage of Different Labels for Utterances from FTF Interviews and AVIs



Note. The proportion differs between the two interview conditions, while for both the categories are highly imbalanced and labels like PR and SD account for little proportion of the data. SS = story situation, STA = story task-action, SR = story results, PS = pseudo-story situation, PTA = pseudo-story task-action, PR = pseudo-story results, VO = values/opinion, SD = self-description, JS = justification, OR = other utterances.

Figure 2

Histogram of the Word Counts and Utterance Lengths in Our Dataset



Note. The distribution leans to the left, indicating a large number of short utterances.

We mainly leveraged the forced alignment toolkit in Massive Multilingual Speech (MMS) (Pratap et al., 2023), an ASR system focused on multilingual scenarios. The toolkit is based on a multilingual Wav2Vec2-based encoder-only model with Connectionist Temporal Classification loss. Such standard acoustic models allowed us to obtain the exact timestamp of phonemes, and the toolkit could then assign timestamps to the words and achieve forced alignments. In particular, the toolkit is able to identify silent frames, which allowed us to jointly perform VAD and forced alignment precisely. Nevertheless, long silence was generally harmful to the model performance. Therefore, we pre-processed the audio with the toolkit, and removed all the long silences (>4 sec), with 1.6 sec safe margin to allow for some error. All the following processes are based on this new version of recordings with silence removed. Using this toolkit we then obtained the final forced alignment and segmented the recordings into audio utterances. Additional margins on the left and right of the segment were added if the pause between words permitted. We found the results generally satisfying.

ASR Transcription

We then produced the ASR transcripts of our recordings for the training and evaluation procedures. Since the ASR system that produces the transcripts is an essential part in our pipeline, we need an ASR system that accurately recognize our French speech data and avoid speaker biases (Hickman et al., 2024). We thus chose the OpenAI Whisper (large-v2) (Radford et al., 2023), which is a sequence-to-sequence ASR model pretrained on an enormous amount of online available speech data from diverse speakers, styles (e.g., accent), and languages. Although it is not suitable for forced alignment, it stands among recent ASR systems with the lowest error rates (Hickman et al., 2024). There were discrepancies in the text normalization scheme (e.g., how the numbers and filler words were transcribed) between Whisper transcripts and ours, which contributed to extra "errors". By running Whisper on the audio segments, we nevertheless obtained a total character error rate of 16.0% between the ASR and human transcripts. This low error rate (close to the 13.9% error rate on Common Voice v3 (Radford et al., 2023)) further verifies the forced alignment results, since the spoken content of each segmented clip will match the transcript only when the segmentation (decided by the starting and ending timestamp of each utterance) is accurate.

Modeling

Model Building

We formulated this interview analysis as a multiclass classification task under supervised ML, based on a pipeline combining multiple pretrained DL models. We followed standard ML practice to partition our data randomly into 8:1:1 training, validation, and testing subsets based on interviews and their conditions (Goodfellow et al., 2016a). Utterances from the same interview were grouped in the same subset, and FTF interviews and AVIs were assigned to the subsets in a balanced ratio. For the classification of utterance types, we focused on predicting the label of each utterance individually. Therefore, the task was simplified as standard supervised learning by fine-tuning pretrained models to

predict the labels of certain shorter text or audio utterances. This task was developed in three stages, *baseline*, *information enrichment* and *context enlargement*. An overview of the pipeline is illustrated in Figure 3.

Baseline assessments

We distinguish two types of modeling, depending on the availability of ground truth transcripts. Human transcript-based models are trained and evaluated using human transcripts, while audio-based models are trained with human transcripts but evaluated using audio data converted to ASR transcripts. We then created three different labeling schemes: The first scheme involves *fine-grained labeling*, which distinguishes the ten human-coded categories (i.e., SS, STA, SR, PS, PTA, PR, VO, JS, SD, and OR). The second scheme, *combining non-narratives*, merges VO, JS, SD, and OR labels into a generalized OR category, reducing the total number of categories into seven. The third labeling scheme, *coarse-grained labeling*, merges the SS/STA/SR into a single S category and PS/PTA/PR into PS, resulting in the most simplified three-way classification.

Information Enrichment

We used two techniques to enrich the training of audio-based models with audio information. The first technique is ASR mixing, which involves mixing samples of human transcripts and ASR transcripts during training. Although Whisper ASR (large-v2) is fairly accurate, the automatic system can produce errors compared to the ground truth human transcripts, which will mislead the model. Even if we do not count the errors, there are still differences in transcribing style (e.g. regarding punctuation). These discrepancies lead to *domain mismatch* of the model inputs between training and evaluation. Such mismatch may reduce performance (Goodfellow et al., 2016b), but can be alleviated by training the model on both the source (human) and the target (ASR) transcripts to better adjust the model to tolerate ASR transcription inputs. The second technique used audio inputs, directly injecting the corresponding audio segment extracted by forced alignment into the model, creating a *joint speech-language model*, so that paraverbal or nonverbal cues missing in the transcripts can be accessed by the model. In this case, we leveraged the pretrained Wav2Vec2 as an audio encoder to extract audio features which are then concatenated with the text inputs as the input sequence to the text model.

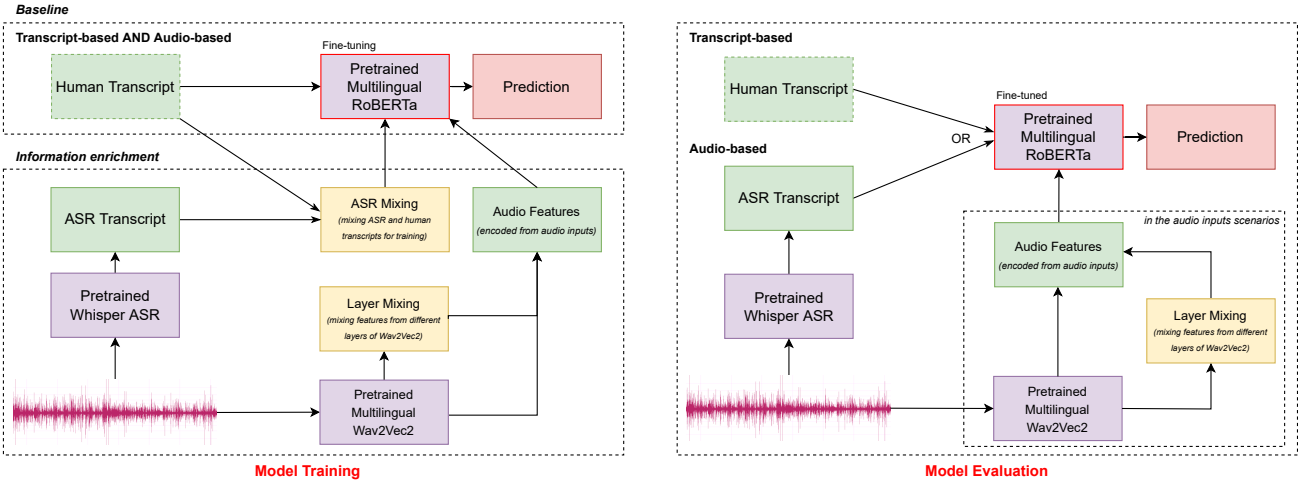
Context Enlargement

To enlarge utterance context, we tested two techniques: coalescence and expansion. Coalescence involves merging adjacent utterances of the same type. Since utterance segmentation is often relatively fine-grained, we can merge adjacent short utterances (< 2 sec) in each interview based on the forced alignment results. To do this, we locate the shortest utterance at each instance using a priority queue and then coalesce it with the neighboring utterance (with < 1-sec gap) of the same type on the left or right side. This allows for the absorption of a short utterance into a longer one or the merger of multiple short utterances into a longer one, resulting in more information being available within each new "utterance" (see Figure 4, b-c). Under the original fine-grained labeling scheme, there are limited adjacent utterances with the same label. Hence, we limit this approach to the coarse-grained labeling case. In this way, the mean length of the utterances can be increased from M = 11.86 words (SD = 5.89) or M = 4.35 sec (SD = 2.66) to M = 35.50 words (SD = 18.70) or M = 12.78 sec (SD = 6.28). Nevertheless, it should be noted that it requires pre-known information about the type of each utterance, which would not be available in real situations.

Expansion is an alternative way to feed the context directly into the model. We provide the model with a longer segment of the interview comprising the target utterance and a short part of context before and after it. We use a pair of special tokens <s> (Begin-of-Sequence) and </s> (End-of-Sequence) to enclose the target utterance within the input to specify the actual target to be classified. With this method, a broader context containing various types of utterances can be included (see Figure 4, d). Specifically, for each target utterance, we included the adjacent utterance before or after the target utterance, starting with the shorter one, so as to cover a context of 15~25 sec with at least 5 sec on each side. FTF interviews involved dialogues between participants and the recruiter, thus also including the interviewer's speech. In this case, for consecutive utterances from one speaker, the exact speaker was indicated to the model (e.g., "CA" for participants). Also, multiple questions were included in FTF interviews and AVIs. When the target utterance is part of the answer to a specific question, the part regarding other questions might not be relevant and should not be included in the context. Therefore, we assumed that in a FTF interview a new question starts whenever the interviewer makes a long statement introducing the next question. As for AVIs, we knew exactly when a new question started as it was on the screen. In both cases, the iterative expansion on either the left or the right side stops when the presumed question boundary is touched, so as to avoid including responses to other questions. In this way, prediction on each utterance may leverage an expanded context of various types of utterances, with M = 19.10 sec (SD = 2.70) or M = 58.83 words (SD = 11.24). Furthermore, this approach allows the model to classify any interview segment with minimal dependence on pre-known information about how the utterances are segmented and labeled.

Figure 3

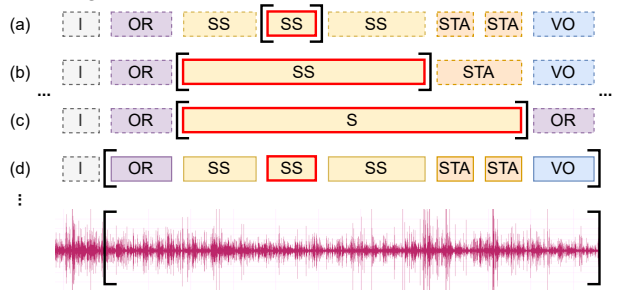
Training and Evaluation Pipeline for Predicting Utterance Types from Interview Recordings



Note. In the transcript-based modeling, the ground-truth human transcripts are used as the sole input for fine-tuning as well as evaluating the pretrained multilingual RoBERTa model; while in the audio-based modeling, human transcripts are still used in training, but transcription by the ASR system, Whisper, becomes the basis of evaluation. Furthermore, in the information enrichment with ASR Mixing, ASR transcripts are also used during training. In addition, the audio inputs may be directly processed by the neural model via Wav2Vec2 as an audio encoder.

Figure 4

Examples of Context Enlargement Techniques with Different Labelling Schemes



Note. (a) Only the target utterance (indicated by the utterance with the red border, with label Story: Situation) is provided; (b) Coalescence: adjacent utterances of identical label are coalesced and classified together; (c) In addition to (b), a more coarse-grained labeling with only three different types (Story/Pseudo-Story/Other Responses) is used, hence the SS and STA utterances can be coalesced; (d) Expansion + Audio inputs: longer context around the target utterance with various labels is provided; in addition, the corresponding audio segment is also provided. I indicates an utterance spoken by the interviewer, which may serve as a boundary between questions.

Model implementation

We use current pretrained language models in our system. For handling text inputs, the model is based on RoBERTa (Liu et al., 2019). RoBERTa is an encoder-only model trained on web texts using masked language modeling. It can be adapted to various downstream language tasks, text classification in our case for example, by fine-tuning it with an extra linear layer stacked upon the encoded feature corresponding to a special token preceding the input text, termed [CLS]. For this task with inputs in French, we used XLM-R, the multilingual version of RoBERTa (Conneau et al., 2020) trained upon data in 100 languages, French included. The model is fine-tuned using Adam optimizer of learning rate 2E-5, with linear warm-up in 500 steps followed by inverse-power learning rate decay, as decided by a grid search. To accommodate the model in a single 16GB V100 Graphing Processing Unit (GPU) while avoiding inefficient padding, we use dynamic batching with around 44 samples per batch in average. We then chose the checkpoint with best validation accuracy, and report the validation and test results (Goodfellow et al., 2016a).

In the case of audio-based modeling, the large-v2 version of Whisper is used as the ASR system to transcribe the recordings for its reported superior performance. Furthermore, instead of relying on ASR transcripts (i.e., indirectly using the audios), we may also allow the model to directly process the audio recordings. To handle audio inputs, we chose the Wav2Vec2 (Baevski et al., 2020) as the basis of our audio encoding module, which is pre-trained on a similar masking-contrastive objective but upon raw audio. In particular, we leveraged its multilingual version XLSR (Conneau et al., 2021) to handle French speech. We observed that the

training failed to converge when we tried to fine-tune both the audio encoder and the stacked text model (i.e., RoBERTa). Therefore, we used the audio encoder as a feature extractor and to additionally concatenate the audio features after the embeddings of the textual inputs as an audio prompt for the text model. The speech models (i.e., Wav2Vec2 and Whisper) are frozen during training, and only the RoBERTa-based text model is updated. We also found out that it is difficult to train the models initialized from the original pretrained RoBERTa. Instead, we fine-tune the model initialized from the text-only response classification built above. Following common BERT-like practice, we leveraged the final layer features of the audio encoder; nevertheless, it has been found that higher layer features represent more semantic cues but lose paralinguistic ones from lower layers. Therefore, we attempted the technique of Layer Mixing with features from different layers averaged using a learnable weight vector, initialized as naive average pooling but updated during the fine-tuning (Pepino et al., 2021), which improves the performance on speech classification tasks.

Results

We present our results after training the models for various techniques. Since the dataset is highly imbalanced (see Figure 1), we report our results using both accuracy and macro-F1 score to measure the model's performance in each setup. The macro-F1 score addresses the imbalance issue by calculating the average of the F1 score on each class. In this way, the performance of the model on less frequent types is better represented. The results on the evaluation (i.e., both validation and test sets) are reported in Table 2.

Baseline Assessments

Results for human transcripts showed satisfying performance in the test set on all three labeling schemes, including the most challenging 10-way fine-grained labeling with 63.65% accuracy and 51.41% macro-F1. The performance was further increased when combining non-narratives (accuracy difference: +2.32%; macro-F1 difference: +5.5%) and even more with the coarse-grained labeling scheme (accuracy difference: +8.69%; macro-F1 difference: +17.82%). The confusion matrix in the human transcript-based model with the fine-grained labeling scenario (see Figure 5) confirmed that the model had trouble distinguishing several categories (e.g., VO vs JS, VO vs OR, STA vs SR). The comparison between human transcript-based and audio-based models showed that the performance of audio-based models was consistently lower across all labeling schemes (accuracy difference range: from -1.73% to -1.86%; macro-F1 difference range: from -0.39% to -1.34%). However, the gap is narrowed as the performance of audio-based models improved considerably from fine-grained labeling to combining non-narratives (accuracy difference: +2.19%; macro-F1 difference: +4.71%) and further to coarse-grained labeling (accuracy difference: +9.54%; macro-F1 difference: +17.98%).

Figure 5

Confusion Matrix Predicted vs True Labels for Human Transcript-Based Fine-Grained Labeling

Note. Distinguishing labels like VO and JS is challenging.

Figure 6

Normalized Attention Weights in Fine-Grained + Audio Input Model Evaluation

	[CLS]	Text	Audio
[CLS]	10.90%	87.72%	1.38%
Text	4.64%	89.18%	6.17%
Audio	0.50%	48.32%	51.18%

Note. Averaged over layers and heads from different part of the inputs on each row to other parts in each column, with [CLS] the classifier token that is used to produce the final prediction. The [CLS] token and the text part puts little attention weight on the audio part, and most attention weights for the [CLS] are placed on the text part of the inputs.

Information Enrichment

Applying ASR mixing in audio-based models during training led to slight improvements in the more challenging fine-grained labeling schemes compared to the audio-based fine-grained labeling model trained on human transcripts only (accuracy difference: +0.33%; macro-F1 difference: +2.66%). But it was less effective with combining the non-narrative (accuracy difference: +0.79%; macro-F1 difference: +0.02%) and coarse-grained labeling (accuracy difference: -0.14%; macro-F1 difference: -0.14%) cases

Table 2

Validation and Test Accuracy/F1 Scores For Baseline Assessment, Information Enrichment and Context Enlargement

	val. acc.	*val. F1.*	*tst. acc.*	*tst. F1.*
HUMAN TRANSCRIPT-BASED				
Fine-grained Labeling	59.83	50.39	63.65	51.41
Combining Non-narratives	62.25	52.61	65.97	56.91
Coarse-grained Labeling	70.16	65.23	72.34	69.23
AUDIO-BASED				
Fine-grained Labeling	56.77	50.24	61.92	50.86
+ ASR Mixing	**57.35**	**51.28**	**62.25**	**53.52** ↑
+ Audio Inputs	**57.06**	**50.43**	61.79	50.76
+ ASR Mixing	57.30	49.40	**62.39**	52.81
+ Layer Mixing	56.87	**51.15**	**61.92**	**51.24**
Combining Non-narratives	59.29	50.25	64.11	55.57
+ ASR Mixing	**59.49**	**51.56**	**64.90**	**55.59**
+ Audio Inputs	**59.44**	**52.07**	**64.76**	**55.69**
+ ASR Mixing	59.24	**52.40** ↑	64.53	55.49
+ Layer Mixing	59.34	51.85	**64.85**	55.61
Coarse-grained Labeling	68.12	63.88	71.46	68.84
+ ASR Mixing	67.25	62.82	71.32	68.70
+ Audio Inputs	68.12	**64.17**	**74.09** ↑	**72.16** ↑
+ ASR Mixing	66.47	62.28	72.38	70.04
+ Layer Mixing	68.12	63.88	71.87	69.32
W/ COALESCENCE				
Coarse-grained Labeling	**69.19**	**63.97**	**72.56**	**69.96**
+ ASR Mixing	**70.27** ↑	**65.21**	**71.87**	**69.58**
+ Audio Inputs	**69.05**	63.20	73.12	70.85 ↑
+ ASR Mixing	**71.35** ↑	**65.63**	**72.56**	**70.28**
W/ EXPANSION				
Fine-grained Labeling	**65.36** ↑	**51.35**	**68.62** ↑	**53.16** ↑
+ ASR Mixing	**65.26** ↑	50.81	**67.60** ↑	**53.71** ↑
+ Audio Inputs	**65.31** ↑	**53.40** ↑	**69.35** ↑	**54.12** ↑
+ ASR Mixing	**65.60** ↑	**50.91**	**67.21** ↑	**53.16** ↑
Combining Non-narratives	**66.91** ↑	**55.58** ↑	**70.57** ↑	54.72
+ ASR Mixing	**66.57** ↑	**54.56** ↑	**71.46** ↑	**58.92** ↑
+ Audio Inputs	**66.62** ↑	**55.30** ↑	**70.14** ↑	54.62
+ ASR Mixing	**66.23** ↑	**54.32** ↑	**71.12** ↑	**59.42** ↑
Coarse-grained Labeling	**72.97** ↑	**68.74** ↑	**77.22** ↑	**74.77** ↑
+ ASR Mixing	**73.80** ↑	**70.35** ↑	**77.45** ↑	**75.39** ↑
+ Audio Inputs	**72.63** ↑	**68.84** ↑	**76.56** ↑	**74.39** ↑
+ ASR Mixing	**74.14** ↑	**70.42** ↑	**77.67** ↑	**75.56** ↑

Note. Results are in %. In each scenario and labeling scheme, the highest results across different informationenrichment techniques are underlined. We also **bold** the results for a technique when it is better compared with the corresponding baselines: In the Audio-based experiments, the results for a specific technique is compared with the one without it (e.g.,+ *Audio Inputs* + *ASR Mixing* is compared with + *Audio Inputs*). In the context enlargement (coalescence/expansion) experiments, one is compared with the Audio-based results with the same enrichment technique but without context enlargement. We further put an ↑ for models with >2% improvements compared to the baseline under the scenario and labeling scheme without any extra enrichment or enlargement. Abbreviations: val. = validation, tst. = test, acc. = accuracy, F1. = macro-F1

compared to their respective audio-based models without ASR mixing. Enriching the model with audio inputs in the training phase improved performance in the more coarse-grained labeling case compared to the coarse-grained labeling audio-based model without audio inputs (accuracy difference: +2.63%; macro-F1 difference: +3.32%). However, it did not improve performance in the fine-grained labeling (accuracy difference: -0.13%; macro-F1 difference: -0.10%) and combining non-narratives labeling (accuracy difference: +0.65%; macro-F1 difference: +0.12%) when compared to the audio-based models without direct audio inputs under their respective labeling scheme. Since the prediction is made upon the features corresponding to the [CLS] token which aggregates the information from the whole inputs using an attention mechanism with an attention weight assigned to each part of the input, we were able to determine the importance of each part of the input for making the prediction by the weight. The normalized attention weights extracted from this model confirmed that the [CLS] token and the transcript part (human or ASR) placed little attention weight on the audio part (see Figure 4). Instead, most attention weights for the classifier token [CLS] were placed on the text part of the input. Moreover, the attention weight on the audio part gradually decreased during the training process, indicating that, for making the prediction, the model learned to put more focus on the texts and less focus on the audio. Therefore, the information learned from the audio by the model was of limited relevance for making the prediction, and a pipelined analysis via ASR transcripts can already achieve performance very close or even better that the the model directly processing audios.

We continued to investigate whether combining the audio inputs with other enrichment approaches (i.e., ASR mixing or layer mixing) leads to better performances. Combining the audio inputs with ASR mixing in the training phase seldom improved the models' performance, regardless of the labeling schemes. When compared to their respective labeling schemes audio-based models with only audio inputs enrichment, the accuracy differences range from -1.71% to 0.60% and macro-F1 difference range from -2.12% to +2.05%. Performance when combining the audio inputs with layer mixing in the training phase never deviated much from the audio-based models with audio inputs under their respective labeling schemes but without layer mixing (accuracy difference range: from -0.51% to 0.32%; macro-F1 difference range: from -1.57% to 0.12%). To summarize, enriching the model with ASR Mixing or Audio Inputs may improve the performance, depending on the labeling scheme, though combining them shows limited help to the model.

Context Enlargement

Coalescence improved performance in all scenarios of coarse-grained labeling compared to the original audio-based models trained without context enlargement (accuracy difference range: from +0.18% to +1.1%; macro-F1difference range: from +0.24% to +1.12%), except for model with audio inputs solely (accuracy difference: -0.97%; macro-F1 difference: -1.31%). The performance of models trained with expansion consistently outperformed almost all the scenarios compared to the original audio-based models (accuracy difference range: from +2.47% to +7.56%; macro-F1 difference range: from -1.07% to +6.69%), with the most significant improvement in fine-grained labeling cases. Specifically, the models reached the best results when either audio inputs, ASR mixing, or both were leveraged, with as high as 69.35%, 71.46%, and 77.67% test accuracy and respective 54.12%, 58,92% and 75.56% macro-F1 on the three labeling schemes respectively.

Discussion

This research explored the possibility to detect applicants' storytelling-related labels from their audio-recorded responses using DL models, which pave the way for developing systems integrating AI-powered automatic feedback on applicants' responses to past-behavior questions. After collecting a unique suite of AVI and FTF interview data and building a specialized data pre-processing pipeline, we designed and experimented on multiple DL systems to address the unique challenges inherent to this specific spoken language understanding task, including amongst other complexity of semantic understanding, errors in ASR, and short utterances with small contexts. The models were built (1) based on audio recordings vs. human transcript-based models as a reference point, using different labeling schemes (i.e., fine-grained labeling, combining non-narrative, coarse-grained labeling), (2) enhanced by information enrichment techniques (i.e., audio inputs, ASR Mixing, and layer mixing), and (3) enhanced by context enlargement techniques (i.e., coalescence or expansion).

We first investigated how human transcript-based and audio-based models compare in their prediction performance. Although it is an unrealistic option to use human transcript-based models, their predictions were more accurate than their audio-based model counterparts, regardless of the labeling schemes. The performance of both models increased as the more coarse-grained labeling scheme was adopted and the number of labels to be predicted was reduced. These findings evidence two essential aspects. First, human transcript-based models outperform the simple audio-based models without extra improvement techniques, which is not surprising due to errors and domain shift from ASR transcripts. Second, reducing the number of labels can help models focus on the discrepancies we care more about (here, storytelling) while avoiding some training difficulties. Moreover, merging some categories (e.g., story vs. pseudo-story vs. other utterance type) may be acceptable for practical

needs. These findings underscore the need to reduce the number of labels to predict and to introduce more sophisticated modeling techniques in the training phase to address the challenges in this task for better performance.

Then, we delved into different information enrichment scenarios, such as ASR mixing, audio inputs, and layer mixing, to see how they affect the audio-based models' performance. Our findings revealed that ASR mixing (i.e., mixing human and ASR transcripts in the training phase) could slightly enhance the models' performance, especially when dealing with numerous imbalanced labels (fine-grained labeling). This can be explained by the fact that mixing data more similar to actual evaluation data may alleviate domain shift (Goodfellow et al., 2016b), which is harmful to the performance. Injecting audio inputs to the model can also lead to some performance improvements in the coarse-grained labeling case, while combining both approaches or with layer mixing fail to reliably improve the results. Notably, using direct audio inputs will considerably increase the model complexity and computational costs, which overshadows the necessity of this technique. The small performance difference between the models with and without direct audio inputs suggests that this particular task of storytelling-related classification can be mostly resolved with semantic understanding from verbal cues without paralinguistic cues. This is not surprising due to the nature of this storytelling analysis task, coded based on the text transcripts. This aligns with prior research evidencing that the text is essential in predicting applicants' personality or interview outcomes (Koutsoumpis et al., 2024). However, this may not hold true for more challenging tasks in interview analysis where paralinguistic or nonverbal cues (e.g.,emotion, intonation) play a role. These findings call for further exploration and refinement of these models and enrichment techniques, while considering their relevance to specific tasks.

Lastly, we addressed the issue of short context in utterances using two techniques to expand the context. Our findings showed that the coalescence technique is beneficial to the model performance, while injecting more context using the expansion technique improved the models' performance even more. However, coalescence requires knowledge about the labels that could not be readily available in practice. Therefore, it may be optimal to use the expansion technique for context enlargement, which outperforms all models, including human transcript-based models. These results support the idea that contextual information is essential for higher performance in predicting the labels. Nevertheless, modern transformer-based models notoriously suffer from quadratic computational costs (Vaswani et al., 2017). This means that the time and space costs grow quadratically with the input length. Introducing longer context, despite allowing more information to be used for prediction, may bring marginal improvement to the performance at a much higher computation cost, as distant context is often irrelevant. Moreover, processing an input of a few thousands of words or minutes of audio may already exhaust the memory of a modern GPU. A balance between the computational resources and the context length can be critical in this area.

Research and Practical Implications

This study holds both research and practical implications. Firstly, it offers new perspectives in job interview training by demonstrating the feasibility of delivering automatically generated feedback tailored to participants' specific responses. This may serve as a complement to traditional time-consuming and expensive coaching (Brosy et al., 2020; Lin-Stephens et al., 2022; Ralston et al., 2003). Such training may thus enhance the efficiency and accessibility of interview preparation for applicants.

Secondly, it sets the stage for using state-of-the-art pretrained language models for automatic semantic analyses of storytelling in job interviews, which is an area yet to be explored. Beyond, the application of DL techniques shows promise by enabling the identification of applicants' more complex verbal behaviors from audio recordings. This broadens the scope of previous research predicting interview aspects (e.g., hirability or personality) based applicants' weak verbal, paraverbal and nonverbal behaviors using various statistical and ML approaches (Chen et al., 2017; Hickman et al., 2022; Holtrop et al., 2022; Naim et al., 2015; Nguyen et al., 2014; Rupasinghe et al., 2016).

Thirdly, this study demonstrates approaches to building a robust pipeline to predict job interview-related tasks from audio inputs directly, involving data pre-processing, forced alignment, ASR, and direct audio modeling. It proposes and tests multiple techniques for creating audio-based models that can effectively managing audio-related issues like disfluent speech, noise, and overlapping voices, and reliably make predictions from audio recordings in practical scenarios.

Limitations and Future Research

Nevertheless, our study faces some limitations. The first limitation of this study resides in its experimental approach. The data was collected in a laboratory setting where powerful microphones were used and placed close to the participants. Thus, we had high-quality audio recordings with minor external noises. In real-life situations, the audio quality may not be as good as in our laboratory settings, which may harm the ASR accuracy and the prediction quality. Future research should investigate how these models perform with lower-quality audio recordings and possibly introduce scenarios that build robust models that focus more on the target's voice instead of noises.

A second limitation is that we are confined to a specific narrow task of storytelling detection in job interview analyses. Although producing stories contributes to applicants'

higher interview performance and hirability, other aspects of stories may also play a role. For example, the positive effect of stories may be limited by the quality of their delivery (e.g., paraverbal and nonverbal behaviors). However, our models did not address such aspects because including more factors would have made the models more complex. Therefore, future work should further explore the more detailed automatic analysis of storytelling, including other aspects of the storytelling behavior.

A third limitation is that the models investigated in this study are still at the exploratory stage and, therefore, not yet adequate to integrate them into coaching platforms for real-time feedback as some software do for nonverbal behaviors (Gebhard et al., 2018; Hoque et al., 2013). With our analysis limited to storytelling classification on each single utterance, we are using textual models with relatively short context length, which is another limitation of our research. It exceeds the capabilities of such models to have a global view on the whole interview process. A critical direction of further research would be developing more advanced models that are capable of handling such long inputs efficiently.

Conclusion

AVIs offer a unique opportunity to provide automated personalized feedback in job interview coaching, directly from the audio-recorded responses, in a time-and cost-efficient manner. While such models are not yet widely available, our developed DL models have shown promise in predicting components of optimal (e.g., storytelling) and less optimal (e.g., pseudo-stories) responses from FTF and AVI audio responses. As such models are in their infancy, future research should develop more advanced models capable of handling longer audio inputs and process aspects of utterances other than semantics.

Acknowledgement

This work was supported by the Swiss National Science Foundation (SNSF) under Grant 10521C_197479.

Declaration Statement

The authors report no conflict of interest.

Data Availability Statement

The data supporting this study findings will be openly available on OSF repository at the time of publication.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? how applicants respond to past behavior questions. *Journal of Business and Psychology*, *29*(4), 593–604. https://doi.org/10.1007/s10869-014-9350-0

Bangerter, A., Mayor, E., Muralidhar, S., Kleinlogel, E. P., Gatica-Perez, D., & Schmid Mast, M. (2023). Automatic identification of storytelling responses to past-behavior interview questions via machine learning. *International Journal of Selection and Assessment*, *31*, 376–387. https://doi.org/10.1111/ijsa.12428

Basch, J. M., Melchers, K. G., Kegelmann, J., & Lieb, L. (2020). Smile for the camera! the role of social presence and impression management in perceptions of technology-mediated interviews. *Journal of Managerial Psychology*, *35*(4), 285–299. https://doi.org/10.1108/JMP-09-2018-0398

Brosy, J., Bangerter, A., & Mayor, E. (2016). Disfluent responses to job interview questions and what they entail. *Discourse Processes*, *53*(5-6), 371–391. https://doi.org/10.1080/0163853X.2016.1150769

Brosy, J., Bangerter, A., & Ribeiro, S. (2020). Encouraging the production of narrative responses to past-behaviour interview questions: Effects of probing and information. *European Journal of Work and Organizational Psychology*, *29*(3), 330–343. https://doi.org/10.1080/1359432X.2019.1704265

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel psychology*, *50*(3), 655–702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., & Hoque, M. E. (2017). Automated video interview judgment on a large-sized corpus collected online. *2017 Seventh International Conference on Affective*

Computing and Intelligent Interaction (ACII), 504–509. https://doi.org/10.1109/ACII.2017.8273646

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, & P. Motlícek (Eds.), *Interspeech 2021, 22nd annual conference of the international speech communication association, brno, czechia, 30 august - 3 september 2021* (pp. 2426–2430). ISCA. https://doi.org/10.21437/INTERSPEECH.2021-329

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale (D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault, Eds.), 8440–8451. https://doi.org/10.18653/V1/2020.ACL-MAIN.747

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/V1/N19-1423

Gebhard, P., Schneeberger, T., André, E., Baur, T., Damian, I., Mehlmann, G., König, C., & Langer, M. (2018). Serious games for training social skills in job interviews. *IEEE Transactions on Games*, *11*(4), 340–351. https://doi.org/10.1109/TG.2018.2808525

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out?: Do machine learning application papers in social computing report where human-labeled training data comes from? In M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Fat* '20: Conference on fairness, accountability, and transparency, barcelona, spain, january 27-30, 2020* (pp. 325–336). ACM. https://doi.org/10.1145/3351095.3372862

Germanier, E., Bangerter, A., Orji, K., Schmid Mast, M., Renier, L. A., He, M., & Garner, P. N. (2024). *Responses to past-behavior questions in face-to-face and asynchronous video interviews: Storytelling, interview performance and criterion validity [manuscript in preparation]* (Unpublished Work).

Germanier, E., Bangerter, A., Renier, L. A., Kleinlogel, E. P., Schmid Mast, M., Jayagopi, D. B., Shubham, K., & Roulin, N. (2023). *Effects of interview medium and culture on applicant storytelling, disfluencies and evaluations in behavioral interviews [unpublished manuscript]* (Unpublished Work).

Goodfellow, I., Bengio, Y., & Courville, A. (2016a). Hyperparameters and validation sets. In *Deep learning*. MIT press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016b). Transfer learning and domain adaptation. In *Deep learning*. MIT press.

Hemamou, L., Felhi, G., Martin, J., & Clavel, C. (2019). Slices of attention in asynchronous video job interviews. *8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*. https://doi.org/10.1109/ACII.2019.8925439

Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J., & Clavel, C. (2019). Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. *AAAI 2019*. https://doi.org/10.1609/aaai.v33i01.3301573

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323. https://doi.org//10.1037/apl0000695

Hickman, L., Langer, M., Saef, R. M., & Tay, L. (2024). Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. https://doi.org/10.1037/apl0001247

Holtrop, D., Oostrom, J. K., van Breda, W. R. J., Koutsoumpis, A., & de Vries, R. E. (2022). Exploring the application of a text-to-personality technique in job interviews. *European Journal of Work and Organizational Psychology*, *31*(6), 799–816.

Hoque, M., Courgeon, M., Martin, J.-C., Mutlu, B., & Picard, R. W. (2013). Mach: My automated conversation coach. *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 697–706. https://doi.org/10.1145/2493432.2493502

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897–913. https://doi.org/10.1037/0021-9010.86.5.897

Huffcutt, A. I., & Murphy, S. A. (2023). Structured interviews: Moving beyond mean validity... *Industrial and Organizational Psychology*, *16*(3), 344–348. https://doi.org/10.1017/iop.2023.42

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Sage Publications.

Kessler, R. (2006). *Competency-based interviews: Master the tough new interview style and give them the answers that will win you the job*. Frankilin Lakes : Career Press.

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., van Breda, W., Zhang, T., & de Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior*, *154*, 108128. https://doi.org/10.1016/j.chb.2023.108128

Langer, M., König, C. J., Gebhard, P., & André, E. (2016). Dear computer, teach me manners: Testing virtual employment interview training. *International Journal of Selection and Assessment*, *24*(4), 312–323. https://doi.org/10.1111/ijsa.12150

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer. https://doi.org/10.1007/978-3-319-98131-4_9

Lin-Stephens, S., Manuguerra, M., Tsai, P.-J., & Athanasou, J. A. (2022). Stories of employability: Improving interview narratives with image-supported past-behaviour storytelling training. *Education + Training*, *64*(5), 577–597. https://doi.org/10.1108/ET-08-2021-0320

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Lukacik, E.-R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1), 1–15. https://doi.org/10.1016/j.hrmr.2020.100789

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In F. Lacerda (Ed.), *Interspeech 2017, 18th annual conference of the international speech communication association, stockholm, sweden, august 20-24, 2017* (pp. 498–502). ISCA. https://doi.org/10.21437/INTERSPEECH.2017-1386

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, *1*, 1–6. https://doi.org/10.1109/FG.2015.7163127

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, *16*(4), 1018–1031. https://doi.org/10.1109/TMM.2014.2307169

OpenAI. (2023). GPT-4 technical report. *CoRR*, *abs/2303.08774*. https://doi.org/10.48550/ARXIV.2303.08774

Orji, K., Bangerter., A., Germanier, E., Renier, L. A., Mast, S., M., M., He, & Garner, P. N. (2024). *Extended responses in asynchronous video interviews: Investigating frequency, content, and interview outcomes [manuscript in preparation]* (Unpublished Work).

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

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, & P. Motlícek (Eds.), *Interspeech 2021, 22nd annual conference of the international speech communication association, brno, czechia, 30 august - 3 september 2021* (pp. 3400–3404). ISCA. https://doi.org/10.21437/INTERSPEECH.2021-703

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., et al. (2023). Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Qin, L., Xie, T., Che, W., & Liu, T. (2021). A survey on spoken language understanding: Recent advances and new frontiers. In Z. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI 2021, virtual event / montreal, canada, 19-27 august 2021* (pp. 4577–4584). ijcai.org. https://doi.org/10.24963/IJCAI.2021/622

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Krause,

E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *International conference on machine learning, ICML 2023, 23-29 july 2023, honolulu, hawaii, USA* (pp. 28492–28518, Vol. 202). PMLR. https://proceedings.mlr.press/v202/radford23a.html

Rahman, W., Mahbub, S., Salekin, A., Hasan, M. K., & Hoque, E. (2021). Hirepreter: A framework for providing fine-grained interpretation for automated job interview analysis. *2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021 - Workshops and Demos*. https://doi.org/10.1109/ACIIW52867.2021.9666201

Ralston, S. M., Kirkwood, W. G., & Burant, P. A. (2003). Helping interviewees tell their stories. *Business Communication Quarterly*, *66*(3), 8–22. https://doi.org/10.1177/108056990306600303

Rasmussen, K. G. (1984). Nonverbal behavior, verbal behavior, resumé credentials, and selection interview outcomes. *Journal of Applied Psychology*, *69*(4), 551–556. https://doi.org/10.1037/0021-9010.69.4.551

Richter, M., König, C. J., Koppermann, C., & Schilling, M. (2016). Displaying fairness while delivering bad news: Testing the effectiveness of organizational bad news training in the layoff context. *Journal of Applied Psychology*, *101*(6), 779–792. https://doi.org/10.1037/apl0000087

Riggio, R. E., & Throckmorton, B. (1988). The relative effects of verbal and nonverbal behavior, appearance, and social skills on evaluations made in hiring interviews 1. *Journal of Applied Social Psychology*, *18*(4), 331–348. https://doi.org/0.1111/j.1559-1816.1988.tb00020.x

Roulin, N., Pham, L. K. A., & Bourdage, J. S. (2023). Ready? camera rolling... action! examining interviewee training and practice opportunities in asynchronous video interviews. *Journal of Vocational Behavior*, *145*, 103912. https://doi.org/10.1016/j.jvb.2023.103912

Roulin, N., Wong, O., Langer, M., & Bourdage, J. S. (2023). Is more always better? how preparation time and re-recording opportunities impact fairness, anxiety, impression management, and performance in asynchronous video interviews. *European Journal of Work and Organizational Psychology*, *32*(3), 333–345. https://doi.org/10.1080/1359432X.2022.2156862

Rupasinghe, A. T., Gunawardena, N. L., Shujan, S., & Atukorale, D. (2016). Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis. *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 288–295. https://doi.org/10.1109/ICTER.2016.7829933

Stevens, C. K., & Kristof, A. L. (1995). Making the right impression: A field study of applicant impression management during job interviews. *Journal of Applied Psychology*, *80*(5), 587–606. https://doi.org/10.1037/0021-9010.80.5.587

Suen, H., Hung, K., & Lin, C. (2019). Tensorflow-based automatic personality recognition used in asynchronous video interviews. *IEEE Access*, *7*. https://doi.org/10.1109/ACCESS.2019.2902863

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html


unine
Université de Neuchâtel