Machine Learning with Applications 9 (2022) 100318

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Distinctive features of nonverbal behavior and mimicry in application interviews through data analysis and machine learning

Sanne Roegiers [a,*], Elias Corneillie [b], Filip Lievens [b], Frederik Anseel [b], Peter Veelaert [a,c], Wilfried Philips [a,c]

[a] Department of Telecommunications and Information Processing, Image Processing and Interpretation, Ghent University, St-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
[b] Department of Human Resource Management and Organisational Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium
[c] imec, Kapeldreef 75, B-3001 Leuven, Belgium

ARTICLE INFO

Keywords:
Mimicry
Nonverbal behavior
Data analysis
Machine-learning
Classification experiment
Feature selection

ABSTRACT

This paper reveals the characteristics and effects of nonverbal behavior and human mimicry in the context of application interviews. It discloses a novel analyzation method for psychological research by utilizing machine learning. In comparison to traditional manual data analysis, machine learning proves to be able to analyze the data more deeply and to discover connections in the data invisible to the human eye. The paper describes an experiment to measure and analyze the reactions of evaluators to job applicants who adopt specific behaviors: mimicry, suppress, immediacy and natural behavior. First, evaluation of the applicant qualifications by the interviewer reveals how behavioral self-management can improve the interviewer's opinion of the candidate. Secondly, the underlying mechanics of mimicry behavior are exposed through analysis of seven nonverbal actions. Manual data analysis determines the frequency features of the actions and answers how often the actions are performed and how often they are mimicked during application interviews. Two of the seven actions are here deemed negligible due too low frequency features. Finally, machine learning is employed to analyze the data in great detail and distinguish the four behavior categories from each other. A Random Forest classifier is able to achieve 55.2% accuracy for predicting the behavior condition of the interviews while human observers reach an accuracy of 32.9%. The feature set for the classifier is reduced to 130 features with the most important features relating to the correlations between the leaning forward actions of the interview participants.

1. Introduction

Mimicry has long been a part of human interaction. People imitate the behavior (e.g. gestures, postures, mannerisms) of others for a variety of reasons and this behavioral mimicry can occur consciously and unconsciously (Chartrand & Dalton, 2009). Consequently, this phenomenon has awoken the interest of researchers in various fields such as social psychology (Ashton-James & Chartrand, 2009), communication (Tickle-Degnen, 2006), neuroscience (van Leeuwen et al., 2009), developmental psychology (Bernieri et al., 1988) and consumer behavior (Van Swol, 2003). In areas as diverse as romantic relationships (Karremans & Verwijmeren, 2008), teaching (LaFrance & Broadbent, 1976), negotiations (Van Swol, 2003), and children–parent interactions (Bernieri et al., 1988), researchers have investigated the meaning and interpretations of body posture, gestures, and eye contact. In work psychology, research has predominantly focused on how nonverbal behavior affects the outcome of employment (selection) interviews (Barrick et al., 2009). The main assumption here is that applicants may be capable of tactically adopting various nonverbal behaviors to shape the interviewer's impression of them.

The employment interview is one of the most common employee selection methods (Macan, 2009). Despite its limitations, it is often even the only method used for assessing the competencies and capabilities of applicants (Levashina et al., 2014). Whereas most earlier employment interview research focused on its psychometric properties (i.e., reliability and validity of interviewers' evaluations) (McDaniel et al., 1994), more recent research has shifted the attention to the interview process and to the interaction between interviewers and applicants (Levashina et al., 2014). One main research area deals with how applicants use verbal and nonverbal self-presentation or impression management tactics to make a favorable impression in interviews (Barrick et al., 2009; Motowidlo & Burnett, 1995). The most frequently studied nonverbal

[*] Corresponding author.
E-mail addresses: Sanne.Roegiers@UGent.be (S. Roegiers), Elias.Corneillie@UGent.be (E. Corneillie), Filip.Lievens@UGent.be (F. Lievens), Frederik.Anseel@UGent.be (F. Anseel), Peter.Veelaert@UGent.be (P. Veelaert), Wilfried.Philips@UGent.be (W. Philips).

https://doi.org/10.1016/j.mlwa.2022.100318
Received 27 August 2021; Received in revised form 10 March 2022; Accepted 2 May 2022
Available online 7 May 2022

behaviors in interviews include eye contact, smiling, and – to a lesser extent – gestures and body posture.

In current employment interview research, interaction dynamics which is a key process inherent in non-verbal behavior has been largely ignored (Krasikova & LeBreton, 2012). Nonverbal behavior is interactive because it is influenced by the dyadic interaction (i.e., between two actors or partners, namely the applicant and the interviewer (Ferris et al., 2009)). An intriguing aspect of this interaction is human mimicry, which refers to an interpersonal phenomenon whereby interacting partners often imitate each others behavior in a controlled or spontaneous fashion, e.g., to gain sympathy or trust. For instance, interviewers anticipating a crucial response during the interview may lean forward to show that they are giving the applicant their full attention. Attentive applicants may reciprocate by also leaning forward, thereby communicating that they understand the importance of the question. Although behavioral mimicry has been studied in social psychology, it is still a blind spot in employment interview research. Recent literature and meta-analytic reviews suggest that not a single study has been published regarding behavioral mimicry in the specific interview context (Chartrand & Dalton, 2009; Chartrand & Lakin, 2013; Chartrand & van Baaren, 2009). Hence, pivotal questions remain unanswered: we do not know whether and how mimicry occurs in employment interviews.

The overall aim of this paper is to answer these pivotal questions by studying human mimicry as a key constituent of nonverbal dynamic and dyadic interaction in employment interviews. This is done by observing and analyzing four different behavioral patterns, one of which is nonverbal mimicry, in mock application interviews.

First of all, the effect of the different behavior types on the applicant ratings is investigated. This sheds light on if and how certain nonverbal behaviors can be tactically adopted to make a more favorable impression in application interviews. We hypothesize that nonverbal mimicry on the part of applicant affects their evaluation in a positive manner and are perceived as a better candidate. Mimicry is assumed to enhance the smoothness of interactions, through enhanced perspective taking, thereby leading to increased liking, understanding, and trust between interacting partners (Chartrand & Bargh, 1999).

Secondly, several types of nonverbal gestures and body posture are observed and analyzed in relation to mimicry. For example, ‘how often is each type of action performed in the interview?’ and ‘how often is this action mimicked?’ are questions that are answered through manual data analysis to investigate how mimicry occurs in employment interviews. The most prominent features (e.g. intensity, duration, variation, trend, etc.) of nonverbal mimicking behavior are here determined and quantified.

Lastly, the manual data analysis is enhanced by machine learning. As humans are limited in their processing power of large amounts of data, certain important features and connections are missed. A deeper level of analysis can be achieved through machine learning. Here, the differences and similarities between the four types of behaviors are scrutinized in more depth. The question of ‘can the four behavior categories be distinguished from each other and to what level?’ is answered. The nonverbal actions relevant to mimicry research and their significant features are revealed.

This paper contributes to nonverbal mimicry behavior research in several ways:

- Nonverbal behavioral mimicry is studied in depth in the previously ignored context of employment interviews.
- A flexible annotation tool is presented that enables detailed manual annotation of nonverbal behavior.
- Mimicry is detected automatically from the annotated behavioral data.
- A novel method using machine learning is proposed to perform deep data analysis of nonverbal behavior. This method is demonstrated and compared to traditional data analysis methods.
- Recommendations are made for further mimicry-related research about relevant nonverbal actions and their important relations.

2. The history of research on behavioral mimicry

This section provides a brief overview of the history of research on behavioral mimicry as context for the research described in this paper.

Behavioral mimicry is usually defined as a person imitating the behavior of another person within a short window of time (typically three to five seconds). The mimicked behavior includes facial expressions, emotional reactions, postures, mannerisms, gestures, micro movements and other motor movements. Mimicry is often automatic as in non-conscious, unintentional and effortless.

Initial research on behavioral mimicry focused primarily on interactions between people who knew each other, e.g. patients and therapists (Charny, 1969), student and teachers (LaFrance & Broadbent, 1976) and parents and children (Bernieri et al., 1988). The occurrence of mimicry was found in every case. Mimicry was also exhibited with a greater effect when the duration of contact increased.

Later studies explored behavioral mimicry between people who did not know each other (Chartrand & Bargh, 1999), showing that, even between strangers, behaviors were mimicked, often unconsciously and without later recall of the participants. This phenomenon was coined the “chameleon effect" as humans use mimicry to blend into social situations much like chameleons use it to blend into their surroundings.

Following research about behavioral mimicry had turned to revealing the facilitators and inhibitors of mimicry behavior. The preexistence of a friendly rapport (Tickle-Degnen, 2006), the desire to associate (Lakin & Chartrand, 2003), prosociality (i.e. an increased interest in relating to others) (Chartrand & Bargh, 1999), similarity of opinions (Van Swol & Drury-Grogan, 2017), positive mood and emotions (van Baaren et al., 2006), and executive functioning (van Leeuwen et al., 2009) were all found to be facilitators and lead to an increase in mimicry behavior. Less studies focused on inhibitors of behavioral mimicry but some inhibitors have been found such as the desire to disassociate (Johnston, 2002) and one’s relationship status (Karremans & Verwijmeren, 2008).

The consequences of mimicry were also uncovered in the more recent research studies. Being mimicked can change the cognitive processing of individuals in different ways: they become more field dependent (van Baaren et al., 2004), notice more similarities (van Baaren et al., 2009), are more likely to conform to stereotypes (Leander et al., 2011), think more convergent (Ashton-James & Chartrand, 2009) and their self-consciousness increases (Guéguen et al., 2011). Mimicry can also increase persuasion and can change the consumer behavior and product preferences (Van Swol, 2003). Being mimicked can boost the ability of self-regulation and self-control (Dalton et al., 2010). Lastly, mimicry also has an individual effect on the embodied cognition such as feeling colder when the mimicry level is higher or lower than expected (Leander et al., 2012).

Mimicry also has social consequences and has been called the “social glue” (Lakin & Chartrand, 2003). As long as it remains unnoticed, mimicry can create liking, empathy, smooth interactions, helping behavior and affiliation between interacting people. Moreover, people seem to unconsciously use mimicry to these effects.

In recent years, research fields such as signal processing, computer vision and machine learning had also become involved in the study of mimicry. A multidisciplinary approach was needed as the complexity of behavioral mimicry became increasingly clear. These algorithms replaced the traditional time consuming and costly behavioral observation methods of manual annotation.

Signal processing and logistic regression was applied to detect nonverbal cues from data obtained with wearable sensors. This method was applied to highlight mimicry related behavioral differences between two leadership styles (Feese et al., 2012).

Computer vision techniques were used to compute accumulated motion images to represent the motion cycle of hand movement during a face-to-face conversation (Sun et al., 2011). Similar cross-correlations of body movements between the conversational partners signified a

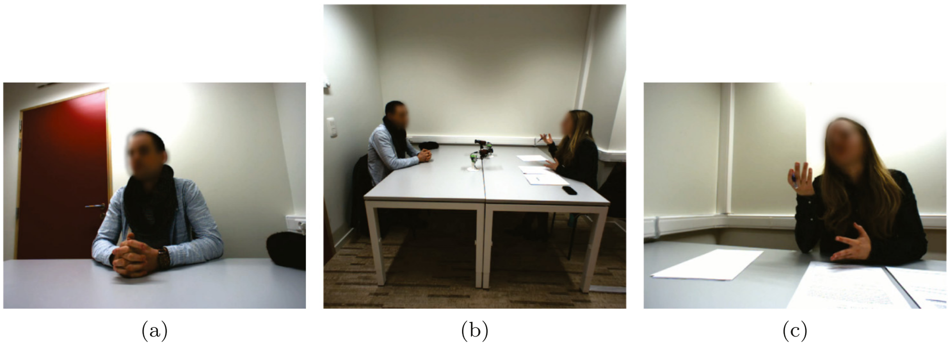


Fig. 1. Camera views of the interview: (a) front view off the applicant, (b) side view of the applicant (left) and the interviewer (right), (c) front view of the interviewer.

high chance of behavioral mimicry in the observed period. In another study, facial keypoint detection combined with facial expression detection was used to research the relationship between mimicry and the perceived quality of the interaction in medical consultations (Wu et al., 2017).

Machine learning techniques were adopted by Bilakhia et al. (2013) to detect audiovisual behavioral mimicry. A long short-term memory neural network utilized facial keypoints to predict mimicry in naturalistic dyadic interactions.

Despite the many number of studies already done on mimicry, behavioral mimicry remains a challenging research area due to its complexity. Researchers continue to delve into behavioral mimicry and illuminate the functions it serves, the underlying mechanisms and the role it plays in building and fostering relationships. However, there are some issues with current mimicry-related research.

As previously mentioned, although behavioral mimicry has been studied in social psychology, it is still a blind spot in employment interview research. Therefore, this paper will focus on nonverbal mimicry behavior in the context of application interviews.

Furthermore, previous mimicry research has primarily focused on only one or two behaviors at a time. The mimicry dynamics between several nonverbal actions and their relevance is largely unexplored. For this reason, a set of seven different nonverbal actions are observed and their dynamics are analyzed for mimicry in this study.

Another issue lies within the traditional method of coding nonverbal behavior. Usually, mimicry is coded by rating the mimicry response on a Likert scale, by counting the number of times the nonverbal actions took place or by estimating the total amount of time the nonverbal action was present. Unfortunately, this research practice hampers progress in studying and understanding human mimicry. Detailed coding of all relevant behavior of two people within an interaction is necessary to study the complexities of mimicry behavior. Awaiting reliable automatic behavior detectors, a software program was developed in this study to make detailed coding of the nonverbal actions manageable. Because of this meticulous coding, a deeper analysis of mimicry behavior is possible and is applied in this study.

3. Methodology

3.1. The study sample

The study experimentally manipulates nonverbal human mimicry of participants in a simulated selection interview context. We opted for a simulated setting, because they have proven successful in studying job application training (Lievens et al., 2015). The sample consists of 251 students of the Ghent University. Students were recruited by e-mail and benefited from the simulated selection process by increasing their experience with selection procedures and by receiving feedback on their performance. The recruited participants followed an elective course which was designed to improve interpersonal skills via an intensive two-and-a-half day training program (including a workshop on employment interviews).

3.2. Procedure and design

Before the interview, the students were briefed. During this briefing, the mock employee interview was proposed in order to develop skills during the course about conversation techniques. After reading this information, the students were divided into four groups and given brief instructions, with differing instructions per group.

The first group (60 students), the "mimic" group, was instructed to mimic the nonverbal behaviors of the interviewer. For example, when the interviewer leans forward, the applicant should do the same.

A second group with 61 students, the "suppress" group, were instructed to maintain a still posture: placing their hands upon their lap/legs and placing both their feet fixed on the ground so they have a relaxed but fixed position. This helped in reducing the natural reflex of mimicking the interviewer.

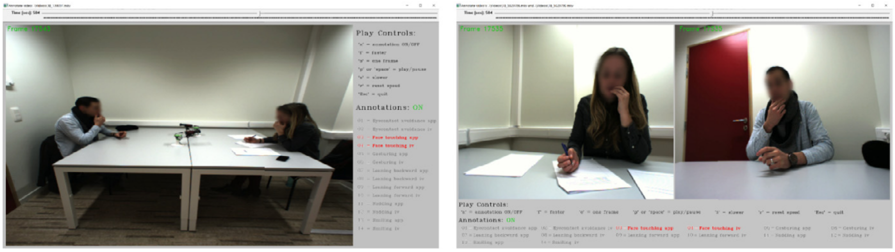
The third group with 67 students, the "immediacy" group, were given instructions to manipulate five nonverbal behaviors during the interview. These behaviors are: holding eye contact with the interviewer, smiling, nodding when the interviewer is speaking, creating a smaller interpersonal distance by leaning forward, supporting speech with gestures (hand movements). Meta-analytic research showed that these nonverbal immediacy behaviors have a positive effect on interviewer ratings of applicants (Barrick et al., 2009; Levashina et al., 2014). This group was therefore included to serve as a good benchmark for the potential effect of mimicry.

These three groups of student applicants were shown a video example of an interview of two trained actors with the 'applicant' actor engaging in the instructed behavior. All applicants were told they could not reveal to the interviewer that they received instructions so that the interviewer remained blind in regards to the manipulations.

The last group (63 students), the "natural" group", did not receive any instructions regarding specific behavior to (not) exhibit. This group only exhibited spontaneous mimicry.

The interviews followed the same procedure as an actual structured employment interview. To increase generalizability to real-world settings, the simulation was kept as realistic as possible (e.g., applicants were asked to send in their resume, interviewers asked questions regarding applicants motivation and interpersonal competencies, etc.). Interviewer questions were counterbalanced to minimize order effects. All interviews were recorded with three cameras (two frontal views and one side view). An example of the camera views are shown in Fig. 1.

After the interview, all applicants were debriefed in another room and were asked to complete a small survey. This survey queried their reactions to the presence of a camera, their motivation to do well on the interview and their perception about the interaction with the interviewer. The interviewer scored the applicant on seven evaluation concepts. Participants received feedback on their interview ratings. Interviewers also described the applicant's salient behaviors (to assess whether they noticed applicant's mimicking behaviors). In a post-interview survey, as a manipulation check, applicants from the first three groups were asked to write down the mimicry instructions they received before and assess the degree to which they conformed to these instructions.



(a) Annotation on the side view of the application interview.

(b) Annotation on the front views of both the applicant and the interviewer.

Fig. 2. Screenshots of the annotation program. The annotation labels for *face touching by applicant*, *face touching by interviewer* and *mimicry* are active in this example.

3.3. Definitions

The terms used in this research are defined in this section.

Mimicry occurs when one of the two conversing people imitates the behavior of the other person within a short window of time. More precisely, once the first person starts a certain behavior or action, mimicry can only ensue if the second person begins to exhibit the same behavior within a time window of 3 s. This means that it is not necessary for both persons to be engaging the same behavior at the exact same time, the person who is mimicked can already have stopped the behavior. The start of the same behavior only has to fall within the 3-second window for mimicry to ensue. There is some debate on the exact onset of mimicry with some studies using a range from 2 to 5 s (Hale & Hamilton, 2016). For the current study, the choice was based on research indicating that mimicry typically occurs within a time frame of 3 s (Tschacher et al., 2013).

The possible mimicked **nonverbal behaviors** are here limited to a set of seven defined actions which include gestures, postures and several other motor movements. The observed actions for mimicry are *smiling*, *nodding*, *avoiding eye contact*, *gesturing*, *face touching*, *leaning forward* and *leaning backward*. A special situation arises in the case of the interviewer: in the course of the interview, the interviewer is taking notes. During the writing, the interviewer is not making eye contact with the applicant, however this is not seen as eye contact avoidance because it is influenced by an outside force.

An **event** is characterized by a start time/frame where the actor starts a nonverbal action, an end time/frame where the actor stops the action, the type of performed action or mimicry and the actor who performs the action.

The **start** of the application interview is defined as the moment the applicant sits down and first looks in the direction of the interviewer. This is seen as the first significant interaction between the two participants. The **end** of the interview is defined as the moment the applicant stands up and leaves the room. The entering and leaving of the room by the applicant is not considered as part of the interview and is not included in the study and annotations.

3.4. Approach to the research questions

As previously mentioned, the first research question of this study is whether nonverbal human mimicry contributes to a better evaluation of the applicant's qualifications in an interview. For this purpose, we compare the interviewers' evaluations of the unmanipulated, natural behavior applicant group to the evaluations of the manipulated mimicry applicant group. The four different groups will also be compared against each other to decide which group performed best overall and per evaluation concept. Finally, we compare the behavior categories of the 20 best evaluated applicants to account for the individual performance instead of group performance.

The second research topic discusses the characteristics of the nonverbal behavior and mimicry (e.g. intensity, duration, variation, trend, etc.). We investigate which features are distinctive for the dissimilar behaviors.

The last posed research question is to find the most prominent features of nonverbal mimicking behavior and to distinguish the behavior type of the applicants through machine learning. A great number of features is extracted from the manual annotations and used as input in a Random Forest classifier. Feature selection is used to reduce the feature set to the most prominent features.

3.5. Annotation

Before the nonverbal behaviors and human mimicry by the participants of the interviews can be examined and analyzed, they have to be detected and annotated. The annotation is here done manually by two human observers who used a behavioral coding scheme, including the definition of mimicry and its nonverbal behaviors. This will be replaced by automatic detection through computer vision techniques in future work.

3.5.1. Annotation process

The annotation of the defined actions is done through an annotation program specifically designed for this purpose. The annotation program allows the coder to choose between two camera views and play the interview video at a desired speed. Two screenshots of the program are shown in Fig. 2. While watching the video, the coder can toggle the annotation label for the expressed nonverbal behavior on and off with simple key presses. A list of annotation labels is visible next to the video with their corresponding number key. These labels are highlighted when active. Rewind, speed up and speed down options are available to increase the accuracy of the annotations. The annotation tool results in a time line of the nonverbal behavior and mimicry where each instance is defined by a start frame and an end frame as shown in Fig. 3.

With this tool, very detailed annotation of the nonverbal behavior by both participants in the interviews can be achieved. However, even with the efficient functionality of the annotation program, this level of annotation detail comes at the cost of a fairly long processing time. To annotate the 251 interviews (approximately 45 h of video), the 2 human coders needed a full month of work (approximately 325 working hours).

3.5.2. Reliability study

A random subset of 20 interview videos were mutually annotated by the coders for a reliability study of the human scorer. The reliability study scores how well the two human coders agree on the annotations. If the annotations between the two coders are too different, the annotations cannot be considered reliable. A reliability study is therefore imperative for the assessment of the annotated nonverbal actions because, without it, there can be no confidence in the annotation nor can any analytic conclusions be drawn from the annotations. A widely used reliability index is the intraclass correlation coefficient (ICC), which reflects the variation between two or more raters who measure the same group of subjects in the case of an interrater reliability study.

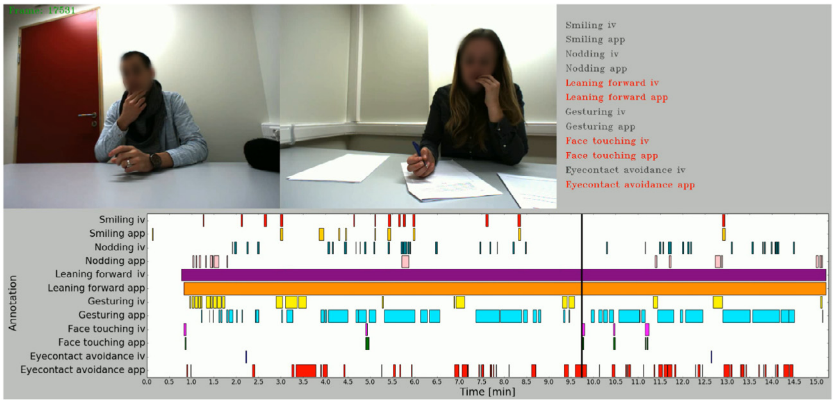


Fig. 3. Annotation timeline with front views of the applicant and interviewer. The colored bars indicate the time the corresponding instance of nonverbal behavior or mimicry is exhibited.

Shrout and Fleiss (1979) defined six different forms of ICC's which were expanded to 10 different forms by McGraw and Wong (1996). The different forms of the ICC are defined by the model (one-way random, two-way random or two-way fixed), the unit of rating (single rater or mean of k raters) and the relationship considered to be important (consistency or absolute agreement).

Following the guidelines by Koo and Li (2016), a specific ICC is selected for rating the reliability of the two annotators for each defined nonverbal action. The selected model is the two-way fixed model as the two raters are fixed and are the only raters of interest. This means that the results only represent the reliability of the two annotators and cannot be generalized to other annotators. The mean result of the two raters is used as the unit of rating. The consistency between the two raters is the considered relationship. Systematic differences between annotators are irrelevant as these represent a systematic small delay in annotation reaction time. As long as the difference in timing for the start and end frame of the action is systematic for each annotator, it does not have any effect on the later analysis of the actions. This form of the ICC is equivalent to Cronbach's alpha and is calculated by:

$$ICC = \frac{BMS - EMS}{BMS} \tag{1}$$

with BMS the between-subjects mean square and EMS the mean square error (Cronbach, 1951; McGraw & Wong, 1996; Shrout & Fleiss, 1979).

The ICC scores are presented in Table 1 for each action event type and each participant of the application interview. The ICC obtained through Formula (1) is only an expected value of the true ICC. This is why the 95% confidence interval is also computed. The level of reliability is determined from this interval. Following the convention suggested by Koo and Li, ICC values above 0.90 indicate excellent reliability, values between 0.75 and 0.90 indicate good reliability, values between 0.50 and 0.75 indicate moderate reliability and lastly, values below 0.50 indicate poor reliability (Koo & Li, 2016). Most of the annotations have a reliability that is good up to excellent. The annotations of the actions *leaning backward* and *nodding by the applicant* have a slightly lower reliability ranging from moderate to excellent. Overall, the expected reliability of the action events annotations is considered to be good to excellent, therefore analytical conclusions can be made with confidence from the annotations.

3.6. Mimicry detection

Our software detects mimicry events automatically from the manually annotated action events. By definition, mimicry only occurs when a participant initializes the same type of behavior within 3 s of the action started by the other participant. Therefore, only the start frame of each action event is needed to detect mimicry.

Table 1
Results of the reliability study for the annotations using the ICC mean rating, consistency, two-way fixed model.

Event type	Actor	ICC	95% confidence interval	Reliability
eye contact avoidance	app	0.965	0.910–0986	excellent
	iv	0.920	0.798–0.968	good–excellent
face touching	app	0.978	0.945–0.991	excellent
	iv	0.981	0.952–0.992	excellent
gesturing	app	0.946	0.865–0.979	good–excellent
	iv	0.909	0.769–0.964	good–excellent
leaning backward	app	0.805	0.508–0.923	moderate–excellent
	iv	0.829	0.569–0.932	moderate–excellent
leaning forward	app	0.931	0.826–0.973	good–excellent
	iv	0.919	0.795–0.968	good–excellent
nodding	app	0.848	0.617–0.940	moderate–excellent
	iv	0.959	0.896–0.984	good–excellent
smiling	app	0.901	0.750–0.961	good–excellent
	iv	0.910	0.773–0.964	good–excellent

Fig. 4 shows an example of mimicry detection. The annotation timeline with the events performed by the applicant and by the interviewer is considered side by side for each type of action. A sliding window equivalent to 3 s starts at the beginning of the timeline and slides to the first start frame of the action events. If a start frame from the other interview participant is detected within this window, mimicry has occurred. If no start frame from the other participant is detected within the sliding window, no mimicry has occurred. The sliding window slides to the next frame and repeats the procedure until the end of the timeline is reached.

3.7. Machine learning

In supervised machine learning, classification is a learning problem that tries to predict the category to which a new observation sample belongs from a predefined set of categories. This is done on the basis of a training set of samples whose category is known. Often, the individual observation samples are summarized into a vector of quantifiable attributes, called features. The features of the training samples are used by the learning algorithm to build a prediction model for the categories of the data.

3.7.1. Classification features

Choosing informative, discriminating and independent features is a crucial step for effective classification. A good feature can distinguish between at least two categories of samples. Unfortunately, finding relevant features is difficult. Therefore, at first, a large set with all sorts of possible relevant features is constructed to classify the different

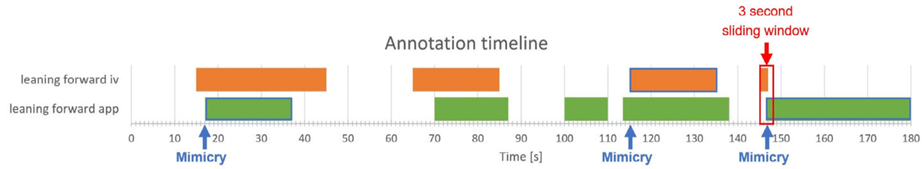


Fig. 4. Example of the mimicry detection on the annotated action events. A sliding window of 3 s slides over the action time line of the applicant and interviewer to the start frame of every event. If the other participant starts the same type of action within the window mimicry is detected.

categories. Later, this initial feature set is reduced by feature selection techniques.

The features are extracted through common analysis techniques from the timelines that were obtained through the manual annotation and the mimicry detection. Each timeline can be seen as a block signal that indicates when a behavioral action or mimicry event is taking place during the interview. The initial set of features extracted from these time signals can be divided into five analysis categories: frequency features (count, frequency, percent), central tendency features (mean, median), dispersion or variation features (range, variance, standard deviation), position features (percentile ranks, quartile ranks) and correlation features (cross correlation, Pearson correlation coefficient).

The cross correlation and the Pearson correlation coefficient are computed according to Eq. (2) and Eq. (3) respectively.

$$(f_{app} \star f_{iv})(\tau) = \sum_{n=-\infty}^{+\infty} f_{app}(n+\tau)\overline{f_{iv}(n)}, \tag{2}$$

with f_{app} and f_{iv} the block signals representing the nonverbal action events off the applicant and interviewer respectively (zero-padded where necessary), τ the time shift between the two signals (ranging from −3 s to 3 s in steps of 0.5 s), and $\overline{f_{iv}(n)}$ the complex conjugate of $f_{iv}(n)$.

$$r(\tau) = \frac{\sum_{n=1}^{N}(f_{app}(n+\tau) - \bar{f}_{app})(f_{iv}(n) - \bar{f}_{iv})}{\sqrt{\sum_{n=1}^{N}(f_{app}(n+\tau) - \bar{f}_{app})^2}\sqrt{\sum_{n=1}^{N}(f_{iv}(n) - \bar{f}_{iv})^2}}, \tag{3}$$

with N the number of frames in the block signals f_{app} and f_{iv}, τ again the time shift between the two signals and $\bar{f} = \frac{1}{N}\sum_{n=1}^{N} f(n)$ the sample mean of the block signal.

Outliers are observations that differ significantly from the other observations in the same category. They are so different from the other data points that they can distort the mean and have a dramatic effect on the correlation between variables. Psychologists often discard outliers from the data to eliminate random error that could be caused by participant or experimental error. Participant expectancies and extraneous variables can affect the final result as can individual differences. However, by removing the outliers the results no longer reflect the actual raw data collected from the experiments. Therefore, in this study, the outliers are not discarded but are isolated into separate features. The outliers are detected based on the InterQuartile Range (IQR) statistic and Tukey's fences (Seo, 2006). The IQR is the distance between the lower quartile Q_1 and the upper quartile $Q3$ of the data points. Tukey's fences determine a range outside which data points are considered outliers. The two fences are located at distance 1.5 IQR below Q_1 and above Q_3. Data points beyond the range $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ are deemed outliers and are isolated from the rest of the data into separate features.

In total, 2188 features are computed for every application interview as the initial feature set. This feature set is later reduced through feature selection techniques.

3.7.2. *Classification model*

In choosing a good classification model, the bias–variance tradeoff is an important consideration (Shalev-Shwartz & Ben-David, 2014). This dilemma concerns the problem of diminishing two sources of error that prevent supervised learning techniques from generalizing beyond their training dataset. The bias error of a classifier stems from underfitting the training data. Important relations between features and class labels are missed during the training phase of the classifier. Conversely, the variance stems from overfitting the training data. Random noise in the training data is then erroneously considered important in the training phase. In general, a classifier that captures the consistency of the training data well and can generalize to unseen data is desired. We want the behavior classification model to achieve a high accuracy for the four behavior types on not only the current dataset of interviews, but also on new interviews. This means a low-bias, low-variance classifier is needed. However, lowering the bias of a classifier, usually implicitly increases the bias and vice versa.

A decision tree classifier is easy to use and can be implemented efficiently (Shalev-Shwartz & Ben-David, 2014). Like any classifier, it takes a set of features and returns a class label, in our case the behavior type, as a prediction. The tree consists of different nodes that are connected by branches. Each decision-node, that is each non-leave node, has exactly two child-nodes in a binary decision tree and represents a split in the feature data. The split is decided by a threshold value for a specific feature. The feature data is continuously split in the tree until leaf-nodes are reached who each represent a single class label.

During the construction of the tree, one feature has to be selected from the entire feature set for the split at each decision-node. Usually the feature with the highest local information gain, or mutual information, is chosen as the split criterion in a greedy heuristic approach. A commonly used criterion to measure the quality of the split is the Gini impurity. It measures the probability of incorrectly classifying a randomly chosen element as if it were randomly labeled according to the class distribution in the dataset. The Gini impurity of the classification outcome for node m is calculated as (Breiman et al., 1984):

$$GI(m) = \sum_{k=0}^{K-1} p_{mk}(1 - p_{mk}), \tag{4}$$

with K the number of classes and p_{mk} the proportion of observations with class k in node m.

When training the decision tree, the best split is chosen by maximizing the Gini Gain, which is calculated by subtracting the weighted impurities of the branches of the node from the original impurity before the split. If node m has two branches resulting in nodes m_{left} and m_{right} with L and R observations in each node respectively, then the Gini Gain of the split in node m is:

$$GG(m) = GI(m) - \frac{L}{L+R}GI(m_{left}) - \frac{R}{L+R}GI(m_{right}) \tag{5}$$

A decision tree is a low-bias high-variance classifier. To decrease the variance of the classifier, multiple decision trees are trained on a different random subset of the population. The resulting classification is decided by a weighted vote from each tree prediction. This method is called the Random Forest Classifier (RF) and is applied in the experiments to classify the behavior type of the applicants (Breiman, 2001). It uses a divide-and-conquer approach to increase the classification performance and is based on the principle that multiple weak classifiers can be combined to create a strong classifier. The structure of this classifier is illustrated in Fig. 5.

The trees in the forest are all trained on a different subset of the training data. However, in practice only a limited amount of training

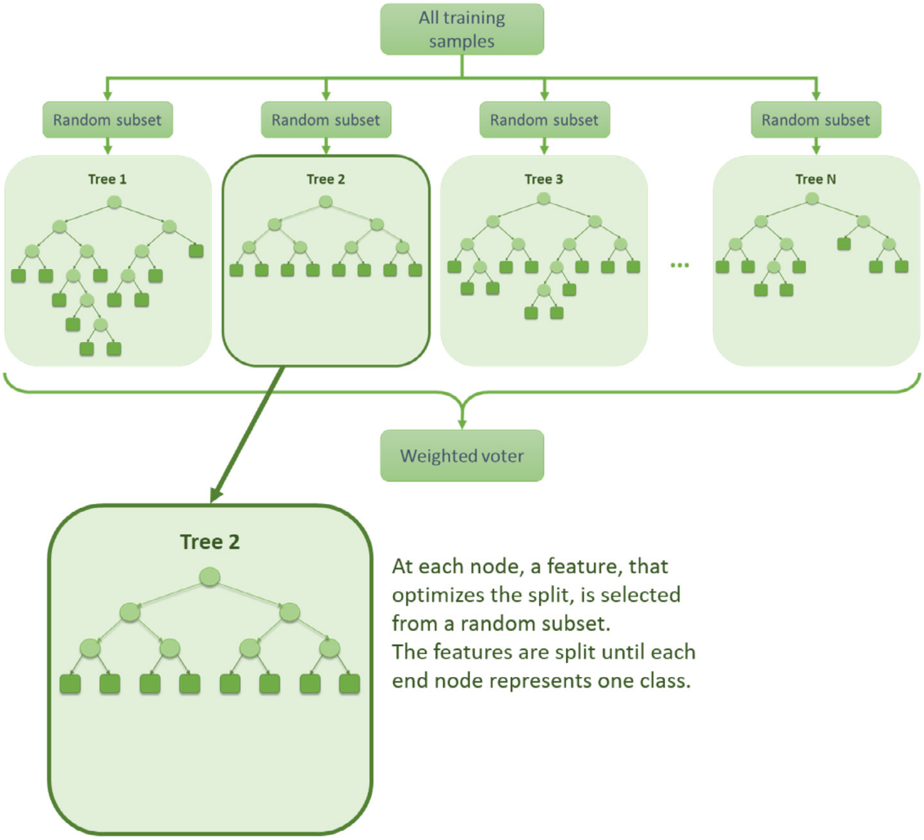

All training samples
Random subset
Random subset
Random subset
Random subset
Tree 1
Tree 2
Tree 3
Tree N
...
Weighted voter
Tree 2
At each node, a feature, that optimizes the split, is selected from a random subset.
The features are split until each end node represents one class.

Fig. 5. Structure of a RF classifier. The classifier is constructed from several binary decision trees on random subsets off the data.

data is available. In our case, the training data is limited to 251 interview samples. A well know method to overcome this problem is sampling with replacement, also called bootstrapping or bagging. This way, a sample can be selected multiple times to train a tree classifier from the forest. As a result, each classifier that is trained on a different bootstrapped sample, will have slightly different decision boundaries. By aggregating the resulting decisions of each of these possibly high-variance classifiers, by means of averaging, a low-variance classifier is obtained. For bagging in RF, the number of trees in the forest is an important parameter to choose. A larger forest is better, but will increase the computing time. Moreover, the classification results will converge and stop getting significantly better beyond a critical number of trees.

The use of bootstrapping to decrease the variance of the final ensemble classifier also has a downside. Because of the duplicates in training data, the bagged trees show a significant correlation between them. Highly correlated trees would therefore make the same errors in similar regions of the feature space. As a consequence, the bias of the resulting classifier increases. To counteract this and to insure diversity in the tree classifiers, randomness is introduced in the splitting criterion: only a randomly selected subset of all features is considered in the split for selecting the feature for the next decision-node.

In summary, a RF, is a low-bias, low-variance ensemble classifier, trained with bagging and random feature selection. It has been proven that RF's are almost invariant to overfitting and are robust against noise. Finally, classification by means of RF has an efficient implementation, since each decision tree can simply be represented by a set of conditional statements.

4. Experiments and results

In the results, the features and characteristics of the three different manipulated behaviors are compared with each other and with those of the unmanipulated natural behavior.

4.1. Evaluation of the applicants by the interviewers

To be able to answer if the posed hypothesis about the positive effect of nonverbal behavioral management on the applicant ratings is true, the evaluations of the applicants are analyzed per behavior type. If the group of applicants who were asked to exhibit mimicry behavior received significantly better evaluations than the group of applicants who did not receive behavioral instructions, then the hypothesis that nonverbal mimicry behavior can affect the interviewer ratings in a positive manner is proven true.

Based on the application interview, the interviewer rates each interviewed applicant on various criteria according to established research norms in psychology (Howard & Ferris, 1996; Oswald et al., 2004). Each criterion is rated on a 7 point Likert scale, ranging from 1 (low) to 7 (high). The resulting evaluation concepts derived from the rated criteria are listed in Table 2.

The boxplot in Fig. 6 shows the results for the evaluation for each group of student applicants. The mean values appear to be higher for the manipulated behavior than the unmanipulated behavior in each evaluation category with one exception in the adaptability evaluation. To confirm this observation and to determine if the mean values are significantly different from each other, the Student's t-test is performed. Table 3 gives the probability $P(T \leq t)$ associated with the Student's two-sample t-test with a two-tailed distribution. The significance level α is set to 5%, meaning that the observed difference between the two sample means is significant if $P(T \leq t) \leq 0.05$. This test states that there is a statistical significant difference in mean evaluation scores between the applicants from the "natural" group and the "mimic" group for qualification, and between the applicants from the "natural" group and "suppress" group for qualification, perceived similarity and affect. In these evaluation categories, the students who manipulated their behavior received better evaluations than the students who showed natural behavior, and thereby (partially) confirming the stated hypothesis that

Table 2
Evaluation concepts with examples used by the interviewer.

Evaluation construct	Example
Continuous learning	Does the applicant tend to improve him or herself in his or her work?
Interpersonal skills	How does the applicant acts with others?
Perseverance	How does the applicant hang on in difficult situations?
Adaptability	How does the applicant adapt towards unforeseen situations?
Qualification	How good is the applicant?
Perceived similarity	Does the interviewer perceive the applicant as similar to himself?
Affect	How does the applicant feel about the applicant (e.g., liking)?

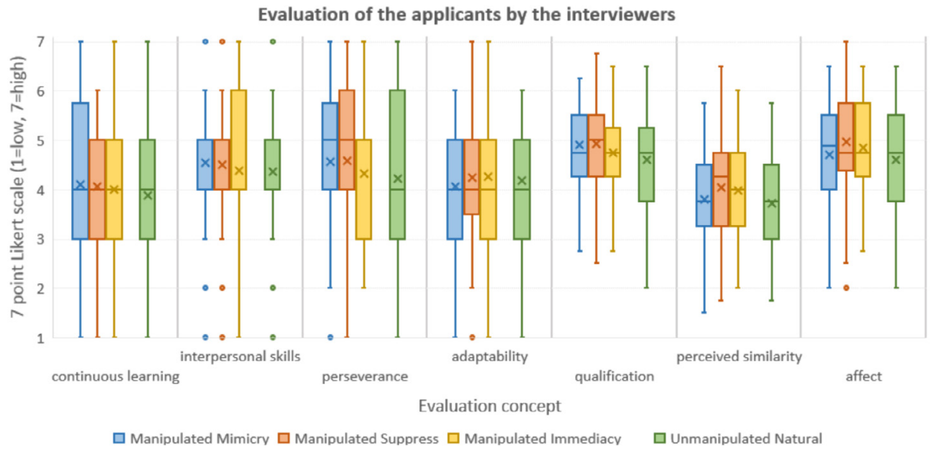


Fig. 6. Boxplot of the evaluation of the applicants by the interviewers. The top of the upper whisker and the bottom of the lower whisker indicate the maximum and minimum value of the sample. The top and bottom of the box are the 75th and 25th percentile respectively. The line through the box represents the median of the sample while the x marker represents the mean. Outliers are indicated by a • marker.

adopting mimicry behavior in an application interview has a positive effect on the evaluation of the qualification of the applicant.

An interesting question is if manipulated behavior has a bigger or smaller effect on top candidates. Table 4 shows the individual evaluation scores for the top 20 applicants. The applicants are ranked here according to the sum of their evaluation scores. From these 20 best evaluated applicants, 40% came from the suppress group, 25% came from the mimicry group, 20% came from the natural group and lastly, 15% of the applicants came from the immediacy group. The fact that most of the applicants in the top 20 come from the suppress group and that the number one applicant is also from this group, corroborates the analysis of the evaluations per group. However, the natural behaving applicants rank rather high on third, fourth, eight and ninth place. This indicates that underlying individual characteristics still outweigh the gains due to manipulated behavior for top candidates.

In conclusion, a strong basis in interview competency is necessary to reach the top rankings in interview evaluations, but the evaluation score can be enhanced by exhibiting strategic nonverbal behaviors such as mimicry behavior or suppressing behavior.

4.2. Manual analysis of the annotations

In this section, analysis of the annotations is used to explore the underlying mechanics of nonverbal mimicry. Prominent features of the nonverbal actions and mimicry are determined and quantified to discover the differences between mimicry, non-mimicry and natural behavior. We distinguish between behavior performed by the applicant and behavior performed by the interviewer. As the interviewers were blind to the mimicry study, it is also interesting to examine if the behavior of the interviewer is influenced by the behavior of the applicant.

4.2.1. Frequency of nonverbal behavior and mimicry

The frequency of the behavior events is an interesting feature for the different interview categories. Some nonverbal actions are more predominant in appearance than other behaviors during the interviews. Aside from being useful to classify the behavior in interviews, the frequency of nonverbal behavior also helps to select the behaviors most suitable for analysis. Event types with a low frequency are better left out of the observation as they can offer only an unreliable amount of data to draw conclusions from.

Fig. 7 shows the average frequency per hour of the action events. The events face touching, leaning forward and leaning backward are very low in frequency in the application interviews. For leaning forward, this is because these events are generally very long in duration. For face touching and leaning backward, it is because this behavior happens only rarely. The most prevalent events are eye contact avoidance by the applicant, nodding by the interviewer and gesturing. From this information it is deducted that face touching and leaning backward are less important behaviors in the study of mimicry because of their minimal occurrence, while eye contact, gesturing, nodding and smiling are important.[1]

Interesting to note is that the naturally occurring behaviors have an overall lower frequency than when they are manipulated in the mimicry or immediacy category, which is in accordance with the instructions to these groups. The instructions of the suppress mimicry group results in a lower frequency of the gesturing and smiling events by the applicant. Even though there is notably less smiling in the manipulated suppress interviews, these applicants received in general a better evaluation by the interviewers.

The influence of the applicant on the behavior of the interviewer is also apparent in this graph. The frequencies of events over the four categories for the interviewer follow the frequencies for the applicant. For example: more smiling by the applicant results in more smiling by the interviewer.

[1] Due to their low frequency, no further significant analysis can be made for the actions face touching and leaning backwards. They are only included for the sake of completion in further analysis.

Table 3
Probability $P(T \leq t)$ associated with the Student's two-sample t-test with a two-tailed distribution. Results in bold green indicate a significant difference between the mean values of the two samples ($\leq \alpha$, with $\alpha = 0.05$).

	Continuous learning	Interpersonal skills	Perseverance	Adaptability	Qualification	Perceived similarity	Affect	Sum
Manipulated Mimicry vs. Unmanipulated Natural	0.25	0.19	0.09	0.31	**0.04**	0.32	0.26	0.15
Manipulated Suppress vs. Unmanipulated Natural	0.24	0.26	0.09	0.42	**0.03**	**0.04**	**0.02**	0.06
Manipulated Immediacy vs. Unmanipulated Natural	0.33	0.40	0.33	0.35	0.17	0.06	0.06	0.15
Manipulated Mimicry vs. Manipulated Suppress	0.50	0.41	0.48	0.24	0.41	0.11	0.09	0.26
Manipulated Mimicry vs. Manipulated Immediacy	0.38	0.27	0.16	0.18	0.17	0.16	0.19	0.48
Manipulated Suppress vs. Manipulated Immediacy	0.38	0.35	0.16	0.42	0.13	0.37	0.29	0.25

Table 4
Evaluation of the top 20 interview applicants.

			Evaluation construct (7-point scale)							Final evaluation
Ranking	Applicant	Category	Continuous learning	Interpersonal skills	Perseverance	Adaptability	Qualification	Perceived similarity	Affect	Sum (max. 49)
1.	129C	Manipulated Suppress	6	7	7	7	6.5	6.5	6.25	46.25
2.	107D	Manipulated Immediacy	7	7	6	7	6.5	5.25	6.25	45
3.	284A	Unmanipulated Natural	7	6	7	6	5.75	5.25	4.75	41.75
4.	93A	Unmanipulated Natural	6	7	5	5	6.5	5.75	6.25	41.5
5.	192C	Manipulated Suppress	4	6	7	5	6.75	5.25	6.75	40.75
6.	123C	Manipulated Suppress	5	6	7	6	6.5	4.5	5.5	40.5
6.	285D	Manipulated Immediacy	5	6	7	7	5.25	5.5	4.75	40.5
8.	151A	Unmanipulated Natural	6	6	6	6	6	4.5	5.75	40.25
9.	207A	Unmanipulated Natural	5	6	5	6	6.25	5.25	6	39.5
9.	25C	Manipulated Suppress	4	6	6	6	6	5.5	6	39.5
11.	138D	Manipulated Immediacy	6	5	5	6	6	5.25	6	39.25
12.	114C	Manipulated Suppress	6	5	5	5	6	5	7	39
13.	58B	Manipulated Mimicry	6	3	6	6	5.75	5.75	6.25	38.75
13.	40C	Manipulated Suppress	6	6	6	4	6.25	5.25	5.25	38.75
15.	197B	Manipulated Mimicry	6	6	3	6	6.25	5.25	6	38.5
15.	32B	Manipulated Mimicry	6	5	4	6	6	5.75	5.75	38.5
15.	226C	Manipulated Suppress	4	6	6	5	6	5.5	6	38.5
15.	33C	Manipulated Suppress	5	5	6	5	5.5	6	6	38.5
19.	137B	Manipulated Mimicry	5	7	5	5	6	4.25	6	38.25
19.	174B	Manipulated Mimicry	6	6	6	4	6.25	4.5	5.5	38.25

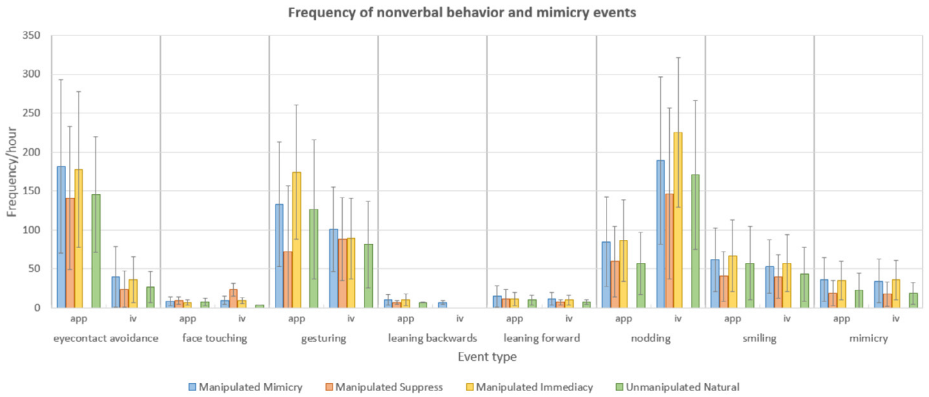


Fig. 7. Frequency per hour of the action events for each interview category. Error bars indicate the standard deviation in regards to the average frequency bars.

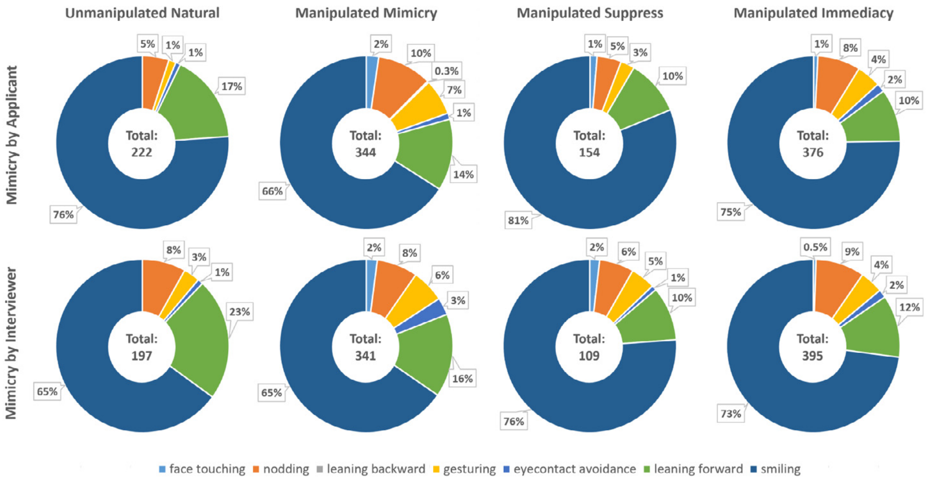


Fig. 8. Total number of mimicry events and the distribution of the mimicry types for the applicant and interviewer of each behavior category.

4.2.2. Mimicry types

Some types of behavior are more easily and more often mimicked than others. The frequency in which an action occurs is not fully correlated with the frequency in which that action is mimicked. The donut charts in Fig. 8 show the composition of the mimicry events and the total number of mimicry occurrences for each actor in each instruction category. Table 5 expresses the portion of the performed action that was mimicked by the other participant in the interview. For example: in the mimicry group, a total of 344 mimicry events by the applicant are observed. Of these 344 mimicry events, 66% is mimicry of smiling (see Fig. 8). When the interviewer is smiling, there is an observed frequency of 45.9% that an applicant from the mimicry group mimics this behavior (see Table 5).

It is clear that the main behavior that is mimicked is smiling even though it is not the action with the highest frequency. The mimicry of smiling is responsible for 65% up to 81% of the all the mimicry events. Smiling is mimicked 23% up to 46% of the time by the other participant. In every category, the applicant mimics smiling more than the interviewer. In the mimicry and immediacy interviews, smiling is mimicked more frequently than in the other interviews. Smiling is thus the easiest and most mimicked action.

The second most mimicked action is leaning forward which pertains to 10% to 23% of the mimicry events. This corresponds with 17% to 46% of the performed leaning forward actions being mimicked. Leaning forward is mimicked most often with natural behavior and least often with suppressed behavior.

In the mimicry and immediacy categories, the applicant and interviewer mimic each other roughly in the same amounts (with a deviation of less than 5%) and with the same distribution of the type of behaviors that are mimicked (with a deviation of less than 2%). These two instruction categories resulted thus in similar mimicry behavior. Even more surprisingly is that the total number of mimicry events by the applicant is almost the same in the immediacy group as in the group that was instructed to mimic the interviewer. The instructions for these two groups resulted thus in very similar behavior in regards to mimicry.

4.3. Deeper analysis through machine learning

Although there are differences to discern between the different behavior categories on the basis of previous manual analysis, it is clear that it will be difficult to determine if the applicant is consciously manipulating his/her behavior. The standard deviation and variance in the data is too great within a category and overlaps to much with the data of other categories. Manual analysis of nonverbal behavior events is insufficient to distinguish the behavior categories from each other. Therefore, machine learning is applied to consider the data in more depth. Machine learning is a great tool to analyze nonverbal behavior in great detail over time. It can find connections in the data that human analyzers miss due to their limited ability to regard and associate large amounts of data over time. For that reason, in this section the four behavior categories (i.e. "mimic", "suppress", "immediacy" and "natural") are classified with the help of machine learning algorithms.

The RF applied in this experiment is trained on the feature dataset previously described. As RF classifiers require few computations, the number of trees and the maximum depth of the trees in the RF are not minimized in this study. The results are obtained with a RF of 2000 trees with maximum 46 features in each tree and the nodes expanded until all leaves are pure. The resulting trees have a depth of minimum 6

Table 5
The observed frequency that a performed action is mimicked by the other participant for each behavior group.

			Manipulated			Unmanipulated
Event type	Actor	Mimicker	Mimicry	Suppress	Immediacy	Natural
eye contact avoidance	app	iv	0.6%	0.1%	0.3%	0.1%
	iv	app	1.4%	0.0%	2.3%	1.2%
face touching	app	iv	41.2%	50.0%	20.0%	0.0%
	iv	app	61.5%	28.6%	60.0%	0.0%
gesturing	app	iv	1.6%	0.8%	0.8%	0.4%
	iv	app	2.3%	0.5%	1.6%	0.3%
leaning backward	app	iv	0.0%	0.0%	0.0%	0.0%
	iv	app	16.7%	0.0%	0.0%	0.0%
leaning forward	app	iv	36.8%	16.7%	35.4%	42.9%
	iv	app	38.3%	22.5%	31.4%	45.7%
nodding	app	iv	3.0%	1.3%	3.6%	2.7%
	iv	app	1.7%	0.5%	1.1%	0.6%
smiling	app	iv	40.2%	23.8%	38.6%	22.5%
	iv	app	45.9%	37.2%	45.4%	37.0%

Table 6
Classification results of the different interview categories.

Classifier	Categories	Accuracy
Human Observers	All	32.9%
Random Forest Classifier	All	55.2% (±1.4%)
Random Forest Classifier	Manipulated, Unmanipulated	54.4% (±4.0%)
Random Forest Classifier	Mimicry, Suppress, Immediacy	64.0% (±4.0%)

splits up to maximum 16 splits and contain between 23 and 42 leaves. This means that most leaves are quickly reached while a small number of leaves hang from long branches.

4.3.1. Behavior classification accuracy

The classification of the different interview categories is evaluated on its accuracy score, meaning the fraction of correct predictions of the classifier on the entire dataset calculated as in Eq. (6).

$$\text{accuracy}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{\sum_{i=0}^{N-1} \hat{w}_i} \sum_{i=0}^{N-1} 1(\hat{y}_i = y_i)\hat{w}_i, \tag{6}$$

$$\hat{w}_i = \frac{1}{\sum_{j=0}^{N-1} 1(y_j = y_i)}, \tag{7}$$

with $\mathbf{y}$ and $\hat{\mathbf{y}}$ the correct classes and the predicted classes of the interviews respectively, N the number of interviews in the dataset and $\hat{w}_i$ the adjusted sample weight to account for dataset imbalance.

When computing the accuracy score of a classifier, it is common practice to use k-fold cross-validation (CV for short). The technique is practiced to avoid overfitting the data and over-evaluating the classifier caused by fitting and testing the classification model on the same data. In k-fold CV, the dataset is split into k smaller sets or folds. The classification model is trained on $k - 1$ folds and validated on the remaining fold. This procedure is repeated k times so each fold has been the validation set once. The reported accuracy is the average of the values computed in each iteration of the loop. The accuracy scores of the RF classifiers presented in this research are the results of 5-fold CV and are disclosed in Table 6.

Our first analysis aims to determine the feasibility of distinguishing the four different types of behavior automatically from the annotated data. This is important because (semi-)automatic behavior categorization is of great help in psychological research and provides deeper insight into the mechanics of nonverbal mimicry behavior.

The RF classifier predicts the category of the interviews from the four possibilities: manipulated mimicry behavior, manipulated suppress behavior, manipulated immediacy behavior or unmanipulated natural behavior. The RF achieves an accuracy of 55.2% (± 1.4%) within a 95%-confidence interval. To evaluate the performance of the RF, the result is compared to the accuracy score of human observers. A balanced subset of 112 interviews were observed by 40 trained psychology students (two to three random interviews per observer). The human observers were able to predict the four interview categories with an accuracy of only 32.9%. To put these numbers in perspective: purely random guessing of the correct category among the 4 possible ones, would lead to an accuracy of 25%. Therefore, human observers perform better than blind guessing, but computer based analysis clearly does a better job.

The above result suggests that computer analysis may be helpful to detect manipulated behavior. In that case, we need to distinguish between two categories: "manipulated" and "unmanipulated". Unfortunately, the accuracy is then only 54.4% (± 4.0%). This is only slightly better than blind guessing, which would achieve an accuracy of 50% on the two classes. This means that it is extremely difficult for the RF to recognize if the interviewer manipulates his behavior with the current data.

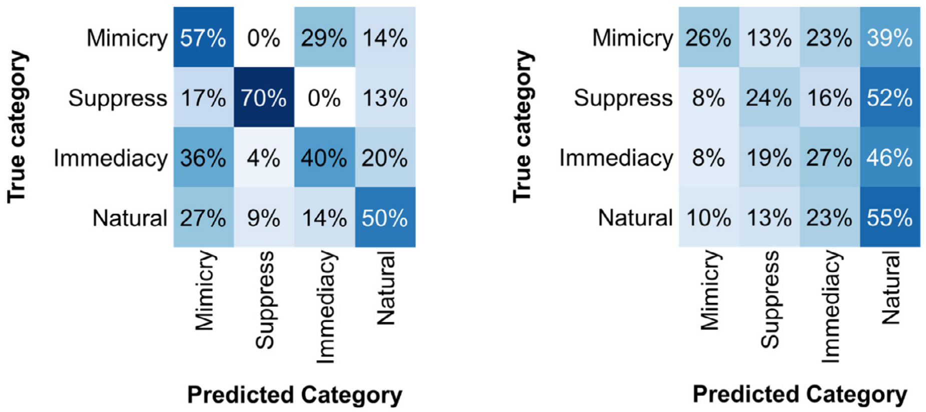
It is much easier for the RF to determine the difference between the three manipulated behaviors mimicry, suppress and immediacy. There, the RF attains an accuracy of 64.0% (± 4.0%) where the blind guessing only manages an accuracy of 33%. The RF can thus predict the type of manipulated behavior with a fairly high accuracy.

4.3.2. Behavior resemblance through the confusion matrix

Evidently the different types of behavior cannot be distinguished from each other with a high certainty. The confusion matrix, or error matrix, of the RF classifier in Fig. 9 is an indication of how much each category resembles the other categories. This matrix presents for each true behavior category the distribution of the predicted labels. This makes it easy to see if the classifier is confusing two (or more) classes by commonly mislabeling one as another.

As the confusion matrix shows, "mimicry" and "immediacy" are very similar in the eyes of the RF classifier and are regularly mislabeled as one-another. Almost 30% of the mimicry interviews are missclassified as the immediacy category and 36% of the immediacy interviews are missclassified as the mimicry category. The "immediacy" category is also the class that is the most difficult to predict. Only 40% of the immediacy interviews are recognized as such. The suppress interviews are the most distinguishable and are correctly classified 70% of the time. This category creates the least amount of confusion with the other categories and is predicted with a high reliability. The natural behavior is also difficult to recognize as only half of these interviews are labeled correctly. It is most often confused with mimicry behavior.

The human observers however, have the tendency to confuse the three manipulated behaviors with natural behavior. Only about a quarter of the manipulated interviews are correctly classified. 39% up to 52% of the manipulated interviews are missclassified as the unmanipulated natural category. Yet, only about half of the natural behavior



(a) RF   (b) Human observers

Fig. 9. Confusion matrix for the classification of the four behavior types in the application interviews.

interviews are correctly labeled as such by the human observer. The human observers thus have an even greater difficulty to differentiate between natural and manipulated behavior.

The biggest confusion for the RF classifier stems from the inclusion of the "immediacy" group. The behavioral actions of this group are significant similar to the actions of the "mimicry" group except for the 3-second time frame necessary for mimicry. If the "immediacy" group were to be omitted from the analysis, the overall performance of the RF classifier would increase to 67% (± 3%) accuracy. In this case, "mimicry" is detected with 54% accuracy, "suppress" with 77% accuracy and "natural" behavior with 72% accuracy.

4.3.3. *Feature selection*

Initially, a large set of features was extracted from the manual annotations because it was unclear at the time which features were relevant for classifying the interview categories. In this section, the feature set is reduced as much as possible without lowering the classification accuracy.

As mentioned before, the Gini impurity is the criterion on which each split in the RF is decided. Each split is based on a decision boundary for one feature and has the maximum Gini Gain possible at that node. When the Gini Gain is accumulated for each time a feature is used as a decision boundary, the importance of that feature in the RF is calculated. After normalization, the relative importance of each feature is known and can be used to reduce the feature set. The least important features are discarded one by one from the feature dataset for as long as the classification accuracy remains within acceptable bounds.

Fig. 10 illustrates the results of the feature selection. The features are ranked from 1 (highest importance) to 2188 (lowest importance) on the x axis. On the right axis, the classification accuracy is plotted in blue for RF's constructed with features of the plotted rank and higher. In light blue, the 95% confidence interval of the accuracy is given. The orange bar chart reflects the relative importance of the ranked features and is plotted on the left axis.

The accuracy of the RF classifier starts to drop rapidly when the number of features are reduced below 130. The decision is therefore made to select the 130 most important features for the feature dataset. At this point, the RF reaches an accuracy of 54.5% (±1.2%) which is still within the 99% boundary of the achieved accuracy with the full feature set.

There are two visible sudden drops in relative importances, one after rank 12 and one after rank 80. The top 12 important features and the top 80 important features are therefore further investigated. Features ranked below 1214 have a relative importance of less than 0.01%. Features ranked below 1813 have no importance and are not used in the construction of the RF model.

Table 7
Distribution of the event types in the top ranked features.

Event type	Top 130 features	Top 80 features	Top 12 features
Leaning forward	72.3%	91.3%	100.0%
Nodding	21.5%	8.8%	0.0%
Gesturing	3.8%	0.0%	0.0%
Smiling	1.5%	0.0%	0.0%
Eye contact avoidance	0.8%	0.0%	0.0%

Table 8
Distribution of the feature types in the top ranked features.

Feature type	Top 130 features	Top 80 features	Top 12 features
Cross Correlation	36.9%	58.8%	0.0%
Pearson Correlation Coefficient	32.3%	32.5%	100.0%
Duration (without outliers)	23.8%	8.8%	0.0%
Frequency	6.2%	0.0%	0.0%
Mimicked by iv/Occurrences by iv	0.8%	0.0%	0.0%

The type of events and the type of features in the selected important features are interesting to note because future research can be limited to the detection of meaningful events and feature types. Table 7 denotes the distribution of the behavior event types that occur in the top 130 of the ranked features while Table 8 denotes the distribution of the feature types. The distribution for the top 80 features and the top 12 features are also included because of the noticeable drop in relative feature importances at these rankings.

Only five of the seven action events are present in this table. Face touching and leaning backward are excluded, which corresponds with the decision in the manual data analysis to omit this data. Leaning forward is the dominant event in the top rankings, distantly followed by nodding, gesturing and an almost negligible amount of smiling and eye contact avoidance. Further similar research into behavioral mimicry should thus focus primarily on leaning forward and nodding as these are prevalent features for classifying the behavior categories.

Five feature types are present in the selected features. Cross correlation features and Pearson correlation coefficients are the most informative in the classification model. They determine a third of the selected features. Pearson correlation coefficients are generally higher ranked than the cross correlation features. Duration features take in less than a quarter of the important features with most of them in the bottom of the rankings. The remainder of the selected features are frequency features and the number of mimicked events relative to the number of event occurrences. It is understandable that correlation features have such a high impact on the classification of mimicry-linked behavior. One of the key characteristics of mimicry is namely

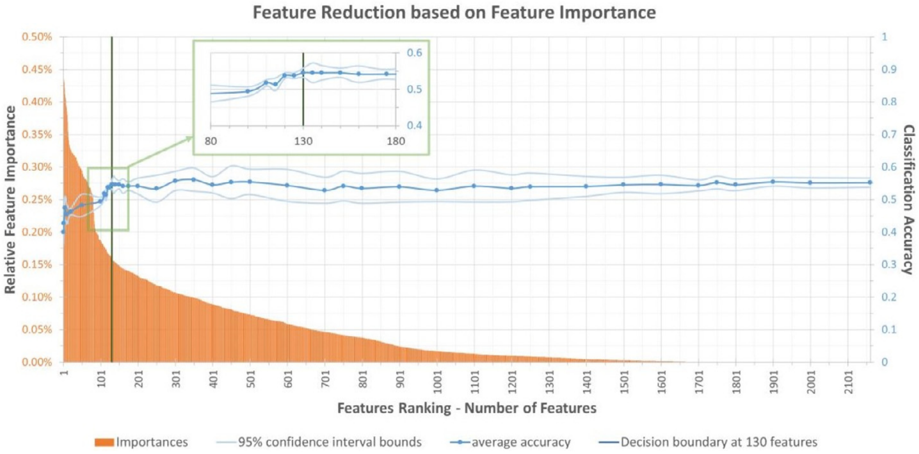


Fig. 10. Feature reduction based on feature importance. Features are ranked from high importance to low importance (orange). Classification accuracy is computed with the number of highest ranking features (blue). The feature dataset is reduced to the 130 highest ranking features without significant loss in accuracy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the timing of the behavior. This timing component is also inherent of the correlation features as one time signal is shifted in time in regards to the other time signal for the calculation.

The 12 most important features are all Pearson correlation coefficients of the action leaning forward at the start of the application interview. This is reflected in the construction of the trees in the Random Forest classification model. These 12 features are, when possible, always the root node of the tree. They decide the first split in the data for the binary tree models. Other root nodes in the decision trees are usually also a correlation feature of leaning forward. The leaning forward interaction between the participants is the most decisive feature for classifying the behavior of the applicant into one of the four categories.

Consequently, the feature dataset can be reduced to 130 features without a significant loss in classification accuracy. The most important features from this set describe the cross correlation, the Pearson correlation coefficient and the duration of the events leaning forward and nodding. Further research should start with these features.

5. Future work

As previously mentioned, the laborious manual annotation process will be replaced by automatic annotation based on computer vision techniques. Human body, hand and facial keypoints will be detected and used as features in action classifiers for the nonverbal behaviors *smiling*, *nodding*, *avoiding eye contact*, *gesturing*, *face touching*, *leaning forward* and *leaning backward*. The manual annotation will serve as the groundtruth to evaluate the performance of the action recognition.

The timing intervals for mimicry can be further analyzed. It will be interesting to find if the response of natural mimicry events are faster or slower than the response of manipulated mimicry events. This experiment can also challenge the definition of mimicry with its 3-second time window.

Another interesting aspect of mimicry is anti-mimicry. With anti-mimicry, instead of exhibiting the same behavior in a 3-second time window, the other participant performs the opposite behavior. For example, when participant A leans forward, participant B reacts by leaning backward. This behavior can be detected and analyzed with the same methods described in this article.

6. Conclusion

This paper described research into the characteristics and underlying mechanics of nonverbal behavior and human mimicry in the context of job application interviews. Three pivotal research paths were explored by observing and analyzing seven behavioral actions (smiling, nodding, avoiding eye contact, gesturing, face touching, leaning forward and leaning backward) during mock application interviews. Applicants were divided into four groups and were instructed to behave in one of four ways: mimicry behavior, suppressing behavior, immediacy behavior and natural behavior.

The first research path challenged the hypothesis that applicants can positively affect the application evaluation by manipulating their behavior. Analysis of the evaluation ratings for the four behavior groups resolved that remaining in a fixed position to suppress mimicry boosted the evaluation ratings. Adopting mimicry behavior also boosted the evaluation scores but to a smaller degree.

The second research path employed manual data analysis of the annotated action events to uncover how mimicry occurs for each behavior type in the interviews. It was discovered that face touching and leaning backwards so rarely occur during the interviews that these events could be discarded. The frequency of the action events were noticeable lower for the suppress behavior in comparison to the other behavior types. It was also observed that the behavior of the applicant influenced the behavior of the interviewer. Lastly, the manual data analysis uncovered that smiling had the greatest frequency of being mimicked and was responsible for the majority of the mimicry events.

The third research path utilized machine learning as a tool to perform a deeper analysis of the obtained data. For each interview, 2188 features were extracted from the annotated and detected mimicry data to classify the interviews into the four different behavior types. A Random Forest was used as the classification model. A classification accuracy of 55.2% (± 1.4%) was reached by the RF compared to an accuracy of 32.9% attained by human observers. The confusion matrix shed light unto why classifying the behavior is difficult. The mimicry and the immediacy group show many similarities and are therefore difficult to distinguish from each other. Finally, the original feature set was reduced to the 130 most important features without a loss in classification accuracy. The most relevant features were found to be the correlation features of the leaning forward actions by the interview participants.

Overall, these results help researchers understand the underlying mechanics of nonverbal behavioral mimicry between applicant and interviewer in an employment interview and form a guideline for further research into this topic.

CRediT authorship contribution statement

Sanne Roegiers: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Elias Corneillie:** Validation, Investigation, Data curation. **Filip**

Lievens: Resources, Supervision. **Frederik Anseel:** Conceptualization, Project administration, Funding acquisition. **Peter Veelaert:** Resources, Writing – review & editing, Supervision. **Wilfried Philips:** Conceptualization, Writing – review & editing, Project administration, Funding acquisition.

References

Ashton-James, C. E., & Chartrand, T. L. (2009). Social cues for creativity: The impact of behavioral mimicry on convergent and divergent thinking. *Journal of Experimental Social Psychology*, *45*(4), 1036–1040.

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.

Bernieri, F., Reznick, J., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology*, *54*, 243–253. http://dx.doi.org/10.1037/0022-3514.54.2.243.

Bilakhia, S., Petridis, S., & Pantic, M. (2013). Audiovisual detection of behavioural mimicry. In *2013 Humaine association conference on affective computing and intelligent interaction* (pp. 123–128). IEEE.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. http://dx.doi.org/10.1023/A:1010950718922.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge.

Charny, E. J. (1969). Psychosomatic manifestations of rapport in psychotherapy. *General Systems Theory and Psychiatry*, 267.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893.

Chartrand, T. L., & Dalton, A. N. (2009). Mimicry: Its ubiquity, importance, and functionality. *Oxford Handbook of Human Action*, 458–483.

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology*, *64*, 285–308.

Chartrand, T. L., & van Baaren, R. (2009). Human mimicry. *Advances in Experimental Social Psychology*, *41*, 219–274.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Dalton, A. N., Chartrand, T. L., & Finkel, E. J. (2010). The schema-driven chameleon: How mimicry affects executive and self-regulatory resources. *Journal of Personality and Social Psychology*, *98*(4), 605.

Feese, S., Arnrich, B., Tröster, G., Meyer, B., & Jonas, K. (2012). Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing* (pp. 520–525). IEEE.

Ferris, G. R., Liden, R. C., Munyon, T. P., Summers, J. K., Basik, K. J., & Buckley, M. R. (2009). Relationships at work: Toward a multidimensional conceptualization of dyadic work relationships. *Journal of Management*, *35*(6), 1379–1403.

Guéguen, N., Martin, A., & Meineri, S. (2011). Mimicry and helping behavior: An evaluation of mimicry on explicit helping request. *The Journal of Social Psychology*, *151*(1), 1–4.

Hale, J., & Hamilton, A. (2016). Cognitive mechanisms for responding to mimicry from others. *Neuroscience and Biobehavioral Reviews*, *63*, http://dx.doi.org/10.1016/j.neubiorev.2016.02.006.

Howard, J., & Ferris, G. (1996). The employment interview context: Social and situational influences on interviewer Decisions1. *Journal of Applied Social Psychology*, *26*, 112–136. http://dx.doi.org/10.1111/j.1559-1816.1996.tb01841.x.

Johnston, L. (2002). Behavioral mimicry and stigmatization. *Social Cognition*, *20*(1), 18–35.

Karremans, J. C., & Verwijmeren, T. (2008). Mimicking attractive opposite-sex others: The role of romantic relationship status. *Personality and Social Psychology Bulletin*, *34*(7), 939–950.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163.

Krasikova, D. V., & LeBreton, J. M. (2012). Just the two of us: Misalignment of theory and methods in examining dyadic phenomena. *Journal of Applied Psychology*, *97*(4), 739.

LaFrance, M., & Broadbent, M. (1976). Group rapport: Posture sharing as a nonverbal indicator. *Group & Organization Studies*, *1*(3), 328–333.

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339.

Leander, N. P., Chartrand, T. L., & Bargh, J. A. (2012). You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological Science*, *23*(7), 772–779.

Leander, N. P., Chartrand, T. L., & Wood, W. (2011). Mind your mannerisms: Behavioral mimicry elicits stereotype conformity. *Journal of Experimental Social Psychology*, *47*(1), 195–201.

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.

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1169.

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

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30.

Motowidlo, S. J., & Burnett, J. R. (1995). Aural and visual sources of validity in structured employment interviews. *Organizational Behavior and Human Decision Processes*, *61*(3), 239–249.

Oswald, F., Schmitt, N., Kim, B., Ramsay, L., & Gillespie, M. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *The Journal of Applied Psychology*, *89*, 187–207. http://dx.doi.org/10.1037/0021-9010.89.2.187.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets*. (Ph.D. thesis), University of Pittsburgh.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press, http://dx.doi.org/10.1017/CBO9781107298019.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Sun, X., Truong, K. P., Pantic, M., & Nijholt, A. (2011). Towards visual and vocal mimicry recognition in human-human interactions. In *2011 IEEE international conference on systems, man, and cybernetics* (pp. 367–373). IEEE.

Tickle-Degnen, L. (2006). Nonverbal behavior and its functions in the ecosystem of rapport. *The Sage Handbook of Nonverbal Communication*, 381–399.

Tschacher, W., Ramseyer, F., & Bergomi, C. (2013). The subjective present and its modulation in clinical contexts. *Timing and Time Perception*, *1*, 239–259. http://dx.doi.org/10.1163/22134468-00002013.

van Baaren, R. B., Fockenberg, D. A., Holland, R. W., Janssen, L., & van Knippenberg, A. (2006). The moody chameleon: The effect of mood on non–conscious mimicry. *Social Cognition*, *24*(4), 426–437.

van Baaren, R. B., Horgan, T. G., Chartrand, T. L., & Dijkmans, M. (2004). The forest, the trees, and the chameleon: Context dependence and mimicry. *Journal of Personality and Social Psychology*, *86*(3), 453.

van Baaren, R., Janssen, L., Chartrand, T. L., & Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*(1528), 2381–2389.

van Leeuwen, M. L., van Baaren, R. B., Martin, D., Dijksterhuis, A., & Bekkering, H. (2009). Executive functioning and imitation: Increasing working memory load facilitates behavioural imitation. *Neuropsychologia*, *47*(14), 3265–3270.

Van Swol, L. M. (2003). The effects of nonverbal mirroring on perceived persuasiveness, agreement with an imitator, and reciprocity in a group discussion. *Communication Research*, *30*(4), 461–480.

Van Swol, L. M., & Drury-Grogan, M. L. (2017). The effects of shared opinions on nonverbal mimicry. *Sage Open*, *7*(2), Article 2158244017707243.

Wu, K., Liu, C., Taylor, S., Atkins, P. W., & Calvo, R. A. (2017). Automatic mimicry detection in medical consultations. In *2017 IEEE life sciences conference* (pp. 55–58). IEEE.