The Expression of Success: Are Thin-Slices of Pre-performance Nonverbal Behavior Prior to Throwing Darts Predictive of Performance in Professional Darts?

Philip Furley¹ · Daniel Memmert¹

Published online: 4 October 2020
© The Author(s) 2020

Abstract

The present research attempted to test how skilled people are at predicting perceptual-motor performance of professional darts players based on short observations of pre-performance nonverbal behavior. In four thin-slices experiments (total $N=490$) we randomly sampled stimulus material from the 2017 World Championships of Darts showing short video recordings of the players immediately before throwing darts. Participants were asked to estimate the points scored for the respective throws. Results across four experiments, all of which were successfully replicated in direct replication attempts, supported the hypothesis that pre-performance nonverbal behavior of professional darts players gives valid information to observers about subsequent performance tendencies. The present research is the first to show that highly skilled individuals seem to display nonverbal cues that observers can pick up to draw inferences about how these individuals are likely to perform.

Keywords Person perception · Social cognition · Nonverbal behavior/communication · Facial expressions · Body language · Perceptual-motor performance · Sports

People are constantly trying to predict how other people are likely to behave. An important source of information in this respect is the momentary nonverbal behavior (e.g., body language, facial expression, etc.) of the target person. Over 60 years of research on *interpersonal accuracy* or *interpersonal sensitivity* has conclusively shown that people in general (despite important individual differences in this skill) are accurate at assessing other peoples' emotions, personality, intentions, motives, and thoughts (Hall et al. 2016; Schmid Mast and Hall 2018). These assessments are important for predicting how another person is likely to behave in the future and give guidance for successfully managing social

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10919-020-00342-2>) contains supplementary material, which is available to authorized users.

✉ Philip Furley
p.furley@dshs-koeln.de

¹ Institute of Training and Computer Science in Sports, German Sport University, Am Sportpark Müngersdorf 6, 50933 Cologne, Germany

interactions. However, limited research has directly tested if people can make accurate immediate behavioral predictions from short observations of nonverbal behavior.

Typically, interpersonal sensitivity is measured with performance tests in which participants see and/or hear other people's behavior (usually in short videos or photographs) and are asked to infer something about the target person. This inference is subsequently compared to a criterion variable. The paradigm that is usually used to determine interpersonal sensitivity at the group level (i.e., people, in general, can make these inferences) has been termed *thin slices paradigm* (Ambady et al. 2000; Ambady and Rosenthal 1992; Carney et al. 2007). Within this paradigm, participants have to, for example, judge what kind of emotion a target person is displaying (i.e., judging a state) or certain personality characteristics (i.e., judging a trait) and these judgments are subsequently compared to self-report measures of emotions and personality (Hall et al. 2008).

Several theoretical accounts predict that humans are efficient at communicating internal states (Bogdan 1997) like emotions (Darwin 1872; Ekman 1992) or social intentions (Fridlund 1994) nonverbally. Most of these theories have been developed to account for facial expressions. The most prominent are arguably *basic emotion theory* (e.g., Ekman 1992; Keltner et al. 2019) and *behavioral ecology theory* (Fridlund 1994; Crivelli and Fridlund 2019). Although the emotions view (basic emotion theory) and the social intentions view (behavioral ecology theory) are rival theories on several accounts and have led to vivid debate (Crivelli and Fridlund 2018; Ekman 2016), they both share the tenet that humans have evolved to be highly sensitive to the nonverbal (facial) displays of others and are (often) able to draw accurate inferences from these displays. This basic tenet is also shared by broader theories of person perception, e.g., the ecological theory of person perception (McArthur and Baron 1983; Zebrowitz and Collins 1997) that assumes that the perceptual abilities of humans have been shaped by their ecological utility. As a result, people have become especially attuned to nonverbal cues (e.g., facial expressions) of other people in order to facilitate adaptive (social) behavior.

Empirical research using the thin slices paradigm has demonstrated that people can make a broad range of accurate inferences from watching short recordings of other people. For example, observer judgments correspond to a high degree with (self-reported) personality characteristics and dispositions of target individuals (Borkenau et al. 2004; Todorov et al. 2005; Willis and Todorov 2006). Further, observers have been shown to accurately infer mental states of perceived individuals (e.g., Baron-Cohen et al. 1996), what the outcome of a perceived conversation is going to be (Curhan and Pentland 2007), how student evaluations of a perceived teacher/professor are going to be (Ambady and Gray 2002), how successful CEOs run a company (Rule and Ambady 2008), how successful salespeople are (Ambady et al. 2006), whether a referee is communicating an ambiguous decision or an unambiguous decision in sports (Furley and Schweizer 2016b). However, we are not aware of research that has used the thin-slices paradigm to test accuracy in predicting immediate perceptual-motor performance after short observations of pre-performance nonverbal behavior.

A problem within thin-slices research has been establishing objective and externally valid criteria against which to evaluate predictions and judgments (Kruglanski 1989). A recent line of thin-slices research has therefore used stimulus materials from televised sports competitions as the outcome of sports competitions are objectively defined by scores or other performance-related statistics (Furley 2019; Furley and Schweizer 2020). This research has for example shown that participants can infer the current score based on

short recordings (or pictures) of athletes' nonverbal behavior during the game (Furley and Schweizer 2014, 2016a).

In summary, humans are constantly displaying nonverbal behaviors—whether they want to or not (DePaulo 1992)—that are associated with certain internal states and therefore inform other people how they are currently feeling and likely to behave. This basic fact has led to the frequently cited aphorism of Paul Watzlawick “one cannot not communicate” (Watzlawick et al. 1967, p. 51). Based on this theorizing, the present research attempted to test if thin-slices of an athlete's nonverbal behavior sampled immediately before performing a certain sports skill can be used by observers to make accurate inferences about subsequent performance tendencies. On a theoretical level, this can be considered an important extension to the thin-slices literature as it would suggest that a person's momentary nonverbal behavior does provide valid information to an observer about how another person is likely to behave/perform in the next moment. While empirical findings have supported that thin-slices of expressive behavior are predictive of behavior/performance at longer timescales (e.g., Ambady et al. 2006; Curhan and Pentland 2007), we are not aware of any research that has shown that a glimpse of nonverbal behavior is predictive of performance immediately following the nonverbal expression. Hence, the present research sought to address this research gap by testing if short recordings of the face and body of professional Darts players immediately before throwing darts could inform observer inferences about performance tendencies of the target players. Further, the present research attempted to remedy some of the shortcomings of previous thin-slices research in sports.

Overview of the Present Research

We attempted to test the central hypothesis that pre-performance nonverbal behavior is associated with subsequent performance in professional darts. One important cue informing observers about performance tendencies of darts players might be located in the eyes, as research on the so-called *Quiet Eye* (Rienhoff et al. 2016; Vickers 2007 for reviews) has consistently shown that performance in perceptual-motor aiming tasks is associated with a calmer (i.e., longer) fixation when aiming to hit a target. More specifically, this research has shown that a better performance when trying to hit a target is associated with longer quiet-eye durations, that is, longer final fixations on the target before initiating the motor action. Given the importance of eyes in human nonverbal communication (see Argyle 1972, 1990; Matsumoto et al. 2013 for reviews), it seems feasible that observers pick up information from the *Quiet Eyes* of darts players and use this information in predicting subsequent performance.

Research has shown that different nonverbal signals are often correlated and collectively serve to inform observers about internal states of other people (Matsumoto et al. 2013). This has led to the theoretical concept of nonverbal response system coherence (Matsumoto and Hwang 2013). In this respect, previous thin-slices research in sports has shown that different nonverbal cues (in an athlete's face, body, and kinematics) informed observers about the current score line in a game. Although we are not aware of any research that has

linked nonverbal cues besides the eyes of athletes to darts performance, it is plausible that there might be several important information sources in the face, body, and kinematics (cf. Furley and Schweizer 2016a) of darts players associated with subsequent performance.

To test the hypothesis that pre-performance nonverbal behavior would be predictive of subsequent performance, we randomly sampled video clips (see footnote¹ for this procedure) of professional Darts players before throwing darts (i.e., up to one frame before a dart was thrown at the darts board). The context of darts has the advantage of providing numerous performances that are entirely under the control of the individual performer (i.e., not, or hardly, dependent on the interaction with teammates or opponents). Further, performance is measured in a fine-grained manner. Therefore, the context of professional darts seems well suited to test the hypothesis if pre-performance nonverbal behavior is predictive of subsequent performance.

In Experiment 1 and Experiment 2, we sampled two different sets of stimulus material showing pre-performance nonverbal behavior in four performance categories (poor performance, medium-to-good performance, good performance, and perfect performance) and subsequently asked participants to estimate the points scored in the respective videos. In Experiment 3, we used the stimulus material from Experiment 2, and only showed the faces of the darts players to find out if the information present in the faces of the darts players was sufficient to infer performance tendencies (e.g., instead of the kinematics of the throwing arm). In Experiment 4, we cut the videos so that the stimulus material only showed one dart (either hitting a triple or missing the triple) instead of three darts as in Experiments 1–3 to find out if pre-performance nonverbal behavior of one dart provided sufficient information to infer performance tendencies of a single performance. Due to the increasing calls for replications in psychological science (Camerer et al. 2018; Open Science Collaboration 2015), we decided to directly replicate every experiment in this series of studies to determine if the findings would be reproducible.

In all experiments, we hypothesized that pre-performance nonverbal behavior would be predictive of performance ratings. More specifically, we expected to find a linear trend of performance ratings contingent with the performance categories: i.e., the better the performance category, the better the mean ratings in these performance categories.

Experiment 1

Participants

We planned to recruit at least 50 participants for the main study, which would provide sufficient power (0.95) to obtain small-to-medium effects ($f=0.2$) in the 1×4 (number

¹ First, we sampled televised video recordings of almost all matches during the 2017 Professional Darts Corporation (PDC) World Darts Championship. These included 29 of 32 matches in round one, all 16 matches in round two, all 8 in round three, all 4 quarterfinals, both semifinals, and the final. Only for 3 matches of round one, video footage was missing. This procedure gave us video footage of 36,168 individual dart scores (12,056 throws) of 61 of the 64 players. Every throw in the four respective performance categories was serially numbered. Then, we created 20 random numbers (in the respective number ranges within four performance categories) via the page random.org in the four performance categories and cut the selected throws (i.e., three individual clips of pre-performance nonverbal behavior of the respective darts in a throw; each video showed the pre-performance nonverbal behaviour of three individual darts in successive order) with the software Adobe Premiere 4.0.

of triples scored) within-subject design (Faul et al. 2007; cf. Schweizer and Furley 2016). We ran an online version and a lab version of Experiment 1. For the online experiment, participants were recruited and tested online. Data collection in the online experiment was terminated after reaching 88 completed data sets. No demographic data were collected in the online study. In the subsequent laboratory study data collection was terminated after reaching 50 complete data sets ($N=50$, 35 male and 14 female (one participant did not answer the gender question); $M_{\text{years}} = 27.1$; $SD=10.87$). On average, the sample reported to have 6.6 h/year ($SD=10.46$) active darts experience and 5.9 h/year ($SD=11.56$) passive darts (i.e., watching darts on TV) experience. All participants gave informed consent via the experimental software.

Procedure

Stimuli Technical details on the sport of darts can be found in the online supplement (see Online Appendix). To avoid the problem of selection bias (e.g., Fiedler 2011), we decided to sample stimulus material of an entire sports event, the Darts World Championships 2017, and randomly select stimuli¹. For Experiment 1, we randomly sampled 20 video clips in four performance categories (poor performance: zero out of three darts within a throw hits a high triple score; medium-to-good performance: one out of three darts hits a high triple score; good performance: two out of three darts hit a high triple score; and perfect performance: three out of three darts hit a high triple score), resulting in 80 video clips. Importantly, we only categorized darts in which the players were aiming for triples, as the final darts in a leg have to hit a specific double field in order to finish the leg, or are thrown to targets to set up a preferred double segment that has to be hit with the final dart.

The videos consisted of a resolution of either 720×1280 px with a frame rate of 30 fps (frames per second) or 360×640 px with a frame rate of 30 fps depending on the type of video files that were available to us. A throwing sequence was determined as valid, if the dart player was shown in a close-up before and during all three darts in the randomly selected throw. Therefore, the starting point of a throwing sequence was selected by the first frame of which a movement of the throwing arm appeared to prepare and initiate the throw (e.g., when a player started to lift the hand of his throwing arm from a resting position, by holding the hand in front of his body). The last frame before the dart left the hand of the player was defined as the ending point of the throwing sequence. Selected throwing sequences were dismissed, if the sequence was either too short (i.e., < 1 s) or the player was not shown during the throw. As a replacement of the dismissed throw the following throw in the corresponding list for each performance category was used (e.g., if throw 1134 was selected in the poor performance category, but was missing a close-up of the pre-performance nonverbal behavior of the second dart, then throw 1135 was included as stimulus material in the study). Figure 1 shows a sample single frame taken from the stimulus material used in the present research. The mean duration of all videos was 3.92 s ($SD=1.13$ s). The videos did not significantly differ in length between the experimental categories ($p=.235$ in 1×4 [number of triples scored] univariate ANOVA on the length of videos).

We controlled for several further variables (with a series of univariate ANOVAs with the within-subject factor performance category) to make sure that the videos in the four experimental categories did not differ significantly in performance prior to the shown stimulus material as this might arguably affect the nonverbal behavior in the videos. No differences were evident in the score line prior to the shown darts

Fig. 1 Sample frame from the stimulus material used in the present research

($p = .608$; points of the player; $p = .344$ points of the opponent). In addition, both the player ($p = .609$) and the opponent ($p = .908$) had scored similar scores with their preceding three darts in all four performance categories. There were also no differences in whether the player ($p = .931$) had won or lost the prior leg across the four experimental categories. Neither was there a differences in how many legs the player had won ($p = .865$) or lost ($p = .220$) prior to the shown stimulus material in the four performance categories. Taken together, there was no indication that the prior performance of the darts player shown in the stimulus material significantly differed and therefore might have had a systematic impact on the nonverbal behavior shown in the stimulus material used in Experiment 1. Finally, the athletes shown in the stimulus material were of comparable strength/performance level across all four experimental categories as no differences were evident in their seeding during this tournament ($p = .865$) or their placing in this tournament ($p = .339$).

We created two versions of Experiment 1. An online version that was administered via the software Unipark and a Desktop Version that was programmed with PsychoPy (Peirce and MacAskill 2018). The respective software randomly sampled 10 videos from the four performance categories and showed the 40 selected video clips in different random orders for every participant. This approach helps to ensure that results do not depend on specific combinations of stimuli. After every video clip, participants were asked to estimate the score of all three darts of the throw for which they had just seen the pre-performance nonverbal behaviors. All videos were presented silently to ensure that ratings were based on nonverbal behavior and not, for example, crowd noise (see Unkelbach and Memmert 2010). Participants gave their ratings by moving a mouse cursor from the middle (90 points) of a 180 points scale to either the left pole (0 points; poorest possible performance) or right pole (180 points; highest possible performance) and clicking the left mouse button to log in their rating for the respective throw. Before commencing the experiment, perceivers filled out a questionnaire

gathering demographic data (only in the lab experiment). Every perceiver was tested individually on a standard 17 inch notebook placed 60 cm away from the perceivers (this information was not available for the online experiment). After completing the testing procedure, participants were informed about the purpose of the experiment.

Results

Online Study A repeated measures ANOVA with the within-subject factor number of triples scored on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category (0-triple [$M=98.09$ points; $SD=29.06$], 1-triple [$M=103.61$ points; $SD=27.77$]; 2-triple [$M=102.57$ points; $SD=29.94$]; 3-triple [$M=107.82$ points; $SD=28.60$]) on the estimated scores of the target players in the online pilot study, $F(3, 261)=7.861$, $p=.00005$, $\eta_p^2=0.074$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.1$. Follow-up polynomial linear contrasts revealed a strong linear relationship between the score estimates and the score categories, $F(1, 87)=21.728$, $p=.00001$, $\eta_p^2=.200$, demonstrating that the score estimates corresponded in a linear manner with the score categories.

Laboratory Study A repeated measures ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category (0-triple [$M=99.91$ points; $SD=20.92$], 1-triple [$M=111.19$ points; $SD=12.36$]; 2-triple [$M=106.04$ points; $SD=17.84$]; 3-triple [$M=107.49$ points; $SD=15.39$]) on the estimated scores of the target players in the laboratory study, $F(3, 147)=7.356$, $p=.0001$, $\eta_p^2=0.131$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.17$. Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 49)=5.273$, $p=.026$, $\eta_p^2=.097$, demonstrating that the score estimates corresponded in a linear manner with the score categories. Although the linear trend analyses showed a significant linear trend, the descriptive pattern of results revealed the highest estimated score in the 1-triple category, and therefore, only provided weak support for our hypothesis.

Pooled Analyses To increase statistical power and obtain more accurate effect size estimates, we further computed a 2 (online vs. lab) $\times$ 4 (number of triples scored) repeated measures ANOVA on the mean estimated scores of the respective triple categories on the pooled data of the online and the lab study. This revealed a significant main effect of triple category (0-triple, 1-triple; 2-triple; 3-triple) on the estimated scores of the target players, $F(3, 408)=12.271$, $p=.0001$, $\eta_p^2=0.083$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.10$. No significant main effect for type of experiment ($p=.434$) nor an interaction between type of experiment by triple category ($p=.097$) emerged.

The descriptive statistics of the pooled analyses are summarized in Fig. 2.

Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 137)=26.229$, $p=.0001$, $\eta_p^2=.161$, demonstrating that the score estimates corresponded in a linear manner with the score categories. Simple contrast analyses revealed significant differences between the 0-triple category and all other categories ($p<.001$), no differences between the 1-triple versus 2-triple ($p=.105$) and 1-triple versus 3-triple ($p=.371$) category, and significant differences between the 3-triple versus 2-triple category ($p=.019$).

Fig. 2 Mean estimated scores in Experiment 1 as a function of triple category in the pooled data of the online study and the laboratory study. Error bars represent standard errors

Discussion

Results in general supported our hypothesis that pre-performance nonverbal behavior in Darts was predictive of performance estimates. In both the online study and the subsequent laboratory study there was a linear relationship between the score category and the mean performance ratings. Further, the simple contrast analyses showed that poor-performance could be reliably distinguished in professional Darts from the other three performance categories. The only other contrast that was significant was between the 2-triple and 3-triple category. The effect sizes of the overall analyses can be considered medium-to-large by convention. In both experiments, the participants seem to have picked up valid cues that they could use to inform them about performance tendencies of the observed professional Darts players.

Experiment 2

Experiment 2 followed a call by Fiedler (2011), who pointed out the necessity of replicating effects found with one set of stimuli with different stimuli to ensure that the phenomenon of interest does not only apply to a highly specific set of stimulus material but generally applies to the phenomenon of interest. Experiment 2 was identical to Experiment 1 only that we randomly sampled an additional subset of 20 videos per experimental category (total of 80 new videos). In addition, the stimuli could include not only triple-20 s (highest triple; and most aimed at triple) this time, but could also contain triple-19, triple-18, and triple-17 as professional darts players sometimes change their target if a previously thrown dart is blocking up large areas of where they are aiming. This was not accounted for in Experiment 1.

Participants

Sample size considerations were identical to Experiment 1. We again ran an online and a lab version of Experiment 2. Participants in the online study were recruited and tested online. No demographic data were collected in the online study. Data collection was terminated after reaching 84 completed data sets. In the subsequent laboratory study data collection was again terminated after reaching 50 complete data sets ($N=50$; 29 male and 21 female; $M_{\text{years}} = 28.0$; $SD=10.46$). On average, the sample reported to have 1.1 h/year ($SD=1.8$) active darts experience and 0.3 h/year ($SD=0.9$) passive darts experience (i.e., watching darts on TV). All participants gave informed consent via the experimental software.

Procedure

Everything was identical to Experiment 1 with the only exception that video stimuli were different. The mean duration of all videos was 3.77 s ($SD=1.32$ s). The videos did not significantly differ in length between the experimental categories ($p=.385$ in univariate ANOVA on the length of videos). We controlled for the same potentially confounding variables (with a series of univariate ANOVAs) as in Experiment 1. No differences were evident in the score line prior to the shown darts ($p=.450$ points of the player; $p=.777$ points of the opponent). In addition, both the player ($p=.977$) and the opponent ($p=.833$) had scored similar scores with their preceding three darts in all four performance categories. There were also no differences in whether the player ($p=.671$) had won or lost the prior leg across the four experimental categories. Neither was there a difference in how many legs the player had won ($p=.441$) or lost ($p=.713$) prior to the shown stimulus material in the four performance categories. As in Experiment 1, there was no indication that the prior performance of the darts player shown in the stimulus material significantly differed, and therefore, might have had a systematic impact on the nonverbal behavior shown in the stimulus material used in Experiment 2. Moreover, the athletes shown in the stimulus material were of comparable strength/performance level across all four experimental categories as no differences were evident in their seeding during this tournament ($p=.337$) or their placing in this tournament ($p=.347$).

Results

Online Study A repeated measures ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category (0-triple [$M=101.88.09$ points; $SD=22.35$], 1-triple [$M=102.74$ points; $SD=21.46$]; 2-triple [$M=109.50$ points; $SD=23.64$]; 3-triple [$M=112.13$ points; $SD=24.71$]) on the estimated scores of the target players in the online study, $F(3, 249)=8.652$, $p=.00002$, $\eta_p^2 = 0.094$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.14$. Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 83)=19.944$, $p=.00002$, $\eta_p^2=.194$, demonstrating that the score estimates corresponded in a linear manner with the score categories.

Laboratory Study A repeated measures ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category

Fig. 3 Mean estimated scores in Experiment 2 as a function of triple category in the pooled data of the online study and the laboratory study. Error bars represent standard errors

(0-triple [$M=95.94$ points; $SD=14.29$], 1-triple [$M=99.30$ points; $SD=12.93$]; 2-triple [$M=106.43$ points; $SD=13.05$]; 3-triple [$M=104.78$ points; $SD=11.63$]) on the estimated scores of the target players in the laboratory study, $F(3, 147)=16.365$, $p<.00001$, $\eta_p^2=0.250$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.21$. Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 49)=44.462$, $p<.00001$, $\eta_p^2=.476$, demonstrating that the score estimates corresponded in a linear manner with the score categories.

Pooled Analyses A 2 (online vs. lab) $\times$ 4 (number of triples scored) mixed ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category (0-triple; 1-triple; 2-triple; 3-triple) on the estimated scores on the pooled data ($F(3, 396)=16.168$, $p=.0001$, $\eta_p^2=0.109$). A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.10$. No significant main effect for type of experiment ($p=.089$), and no interaction between type of experiment by triple category ($p=.538$) emerged.

The descriptive statistics of the pooled analyses are summarized in Fig. 3.

Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 132)=37.792$, $p=.0001$, $\eta_p^2=.223$, demonstrating that the score estimates corresponded in a linear manner with the score categories. Simple contrast analyses revealed significant differences between the 0-triple category versus the 2-triple ($p=.0001$) and 3-triple category ($p=.0001$), but not the 0-triple and 1-triple category ($p=.290$). Further, between the 1-triple and 2-triple ($p=.0001$) and the 1-triple and 3-triple category ($p=.0001$), but not the 2-triple and 3-triple category ($p=.496$).

Discussion

Both versions of Experiment 2 replicated the findings of Experiment 1 and therefore provide further support for our hypothesis that pre-performance nonverbal behavior in

darts is predictive of performance estimates. There was again a linear relationship between the score category and the mean performance ratings. This time, contrast analyses did not show a significant difference between the 0-triple and 1-triple category, but only between the 0-triple and 2-triple and 3-triple category. Further, significant differences were evident between the 1-triple and 2-triple and the 1-triple and 3-triple category. The effect sizes of the overall analyses can be considered medium-to-large by convention.

Experiment 3

The face can be considered the single most prominent nonverbal channel as it is the most intricate (Matsumoto and Hwang 2013, p. 15): “It is the most complex signaling system in our body. [...] And arguably it is the seat of the greatest amount of information that is conveyed nonverbally. That’s why we have ‘face-to-face’ interactions.” Similarly, Cozolino (2006, p. 154) states “the faces of others may be the single most important source of information in our world”. These quotes highlight the critical role of the human face in broadcasting internal states to the social world. Therefore, it seems feasible that participants in Experiments 1 and 2, and observers in sports more generally, primarily use facial signals to infer the score differences (e.g., gaze behavior, i.e., the “Quiet Eyes” of athletes, Vickers 2007). In this respect, we attempted to rule out that participants used kinematic information in the throwing arm (which was visible in Experiments 1 and 2) to inform their ratings. Given the importance of facial information for humans and the *Quiet Eye* reasoning stated above, we expected to find similar linear relationships between score categories and performance ratings as in Experiments 1 and 2 when only facial information was available to observers. In addition, we attempted to investigate whether domain-specific darts knowledge and/or experience moderated the effects found in Experiment 1 and 2.

Participants

Sample size considerations were identical to Experiment 1 and 2. This time we did not decide to run both an online study and a laboratory study, but instead to identically replicate the study. We first ran Study A and after obtaining the results, we ran Study B. In both studies, data collection was terminated after reaching 50 complete data sets (Study A; $n=50$; 30 male and 20 female; $M_{years}=27.35$; $SD=7.20$; replication Study B; $n=50$; 47 male and 3 female; $M_{years}=24.60$; $SD=7.48$). On average, the sample of Study A reported to have been playing darts for 2.91 years ($SD=3.74$); while replication Study B reported playing for 1.61 years; ($SD=2.01$). Study A reported watching 0.80 ($SD=1.85$) darts tournaments per year (0.35, $SD=0.87$ for replication Study B). 14 participants in Study A reported to have no darts experience (17 in replication Study B) and 33 reported to play recreationally (identical in replication Study B), and 3 played at an organized level (0 in replication Study B). All participants gave informed consent via the experimental software.

Procedure

Everything was identical to Experiment 2 with the only exception that we digitally modified the video stimuli of Experiment 2 so that everything else was blackened except for the

Fig. 4 Mean estimated scores in Experiment 3 as a function of triple category in the pooled data of Study A and replication Study B. Error bars represent standard errors

faces of the darts players. Further we collected further variables assessing domain-specific darts knowledge and experience (described in the “[Participants](#)” section) to analyze if these variables moderated the main effect.

Results

Study A A repeated measures ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category (0-triple [$M=105.92$ points; $SD=11.92$], 1-triple [$M=106.39$ points; $SD=10.57$]; 2-triple [$M=116.17$ points; $SD=11.90$]; 3-triple [$M=115.98$ points; $SD=11.56$]) on the estimated scores of the target players in Study A, $F(3, 147)=24.021$, $p<.00001$, $\eta_p^2 = 0.329$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.21$. Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 49)=63.472$, $p<.0001$, $\eta_p^2=.564$, demonstrating that the score estimates corresponded in a linear manner with the score categories.

Replication Study B A repeated measures ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category (0-triple [$M=105.37$ points; $SD=13.85$], 1-triple [$M=102.81.30$ points; $SD=15.00$]; 2-triple [$M=115.76$ points; $SD=13.12$]; 3-triple [$M=118.76$ points; $SD=11.75$]) on the estimated scores of the target players in the replication Study B, $F(3, 147)=28.986$, $p<.00001$, $\eta_p^2 = 0.372$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.23$. Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 49)=87.362$, $p<.00001$, $\eta_p^2=.641$, demonstrating that the score estimates corresponded in a linear manner with the score categories.

Pooled Analyses A 2 (Study A and Study B) $\times$ 4 (number of triples scored) mixed ANOVA on the mean estimated scores of the respective triple categories revealed a significant main effect of triple category on the estimated scores of the target players on the

pooled data, $F(3, 294) = 52.077$, $p = .0001$, $\eta_p^2 = 0.347$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f = 0.12$. No significant main effect for type of experiment ($p = .823$) nor an interaction between type of experiment by triple category ($p = .120$) emerged.

The descriptive statistics of the pooled analyses are summarized in Fig. 4.

Follow-up polynomial linear contrasts revealed a linear relationship between the score estimates and the score categories, $F(1, 98) = 150.786$, $p = .0001$, $\eta_p^2 = .606$, demonstrating that the score estimates corresponded in a linear manner with the score categories. Simple contrast analyses revealed significant differences between the 0-triple category versus the 2-triple ($p = .0001$) and 3-triple category ($p = .0001$), but not the 0-triple and 1-triple category ($p = .429$). Further, between the 1-triple and 2-triple ($p = .0001$) and the 1-triple and 3-triple category ($p = .0001$), but not the 2-triple and 3-triple category ($p = .290$).

Individual Difference Analyses For Experiment 3, we decided to run additional individual difference analyses across both samples. None of the indicators of darts experience significantly moderated the results (when entered as a covariate). The categorical variables gender and darts level also did not significantly moderate the pattern of results when entered as a between-subject factor in the main analyses.

Discussion

Both Study A and replication Study B of Experiment 3 replicated the findings of Experiment 2 with only facial information displayed. Hence, only facial information seems sufficient to draw accurate inferences about performance tendencies based on pre-performance nonverbal behavior. There was again a linear relationship between the score categories and the mean performance ratings. However, the linear trend was not evident between all performance categories as both the poor and medium performance categories and the good and perfect performance categories were estimated in a comparable manner. This was evident in the contrast analyses that revealed a similar pattern as in Experiment 2. The effect sizes of the overall analyses can be considered large by convention. An important addition of Experiment 3 was further showing that the effect was independent of domain-specific experience and gender of the participants.

Experiment 4

Experiment 4 attempted to test if pre-performance nonverbal behavior of single darts (instead of three darts in one throw) were sufficient to give valid information to observers about the performance. An alternative explanation for the results in Experiments 1–3 might arguably be that prior performance of a dart within a throw feeds back and influences the pre-performance nonverbal behavior of the next throw (e.g., success in a first dart of a throw might show in the nonverbal behavior of the next or the next two darts of a throw). If participants in Experiments 1–3 always saw all three darts of a throw in the same successive order that they were thrown, then it might be that the nonverbal behavior of the first dart of a throw would not be predictive of subsequent performance estimates. To test this, we used the extreme categories (poor performance: 0-triples; and perfect performance: 3-triples) from Experiment 2 and cut the videos so that the stimulus material only showed one dart (either hitting a triple or missing the triple) instead of three darts as

in Experiments 1–3. We expected that performance should be estimated higher for all the single triple-darts as compared to the no-triple darts.

Participants

Based on the first three experiments we again planned to recruit at least 50 participants for the study. We again ran a Study A and a replication Study B. We first ran Study A and after obtaining the results, we ran Study B. In Study A, we terminated data collection after reaching 57 participants (39 male and 17 female; $M_{\text{years}} = 34.05$; $SD = 14.44$). On average the sample reported to have been playing darts for 7.38 ($SD = 9.12$) years and to watch 1.04 ($SD = 1.45$) darts tournaments per year. 32 participants reported to have no darts experience, 21 reported to play occasionally, and 7 reported to play recreationally, and 1 played at an organized level. In replication Study B we terminated data collection after reaching 64 participants. Three participants had to be excluded from data analyses as they did not perform the experiment and logged in their answers arbitrarily. This showed by never moving the mouse cursor and always logging in the same answer (either 30 [the middle of the scale], 0 [the left pole of the scale], or 60 [the right pole of the scale]). This left as with a final sample of 61 participants (31 male and 30 female; $M_{\text{years}} = 27.35$; $SD = 7.20$). On average the sample reported to have been playing darts for 2.45 ($SD = 6.37$) years and to watch 1.10 ($SD = 2.41$) darts tournaments per year. 18 participants reported to have no darts experience, 29 reported to play occasionally, and 10 reported to play recreationally. All participants gave informed consent via the experimental software.

Procedure

In Experiment 4, we cut the video stimuli from the 0-triple category and the 3-triple category used in Experiment 2 into individual darts. Altogether, this resulted in a total of 120 video clips (20-first-dart-miss, 20-second-dart-miss, 20-third-dart-miss, 20-first-dart-hit, 20-second-dart-hit, 20-third-dart-hit). This time every video stimuli only showed one dart and participants were asked to estimate the score of every individual dart on a continuous scale ranging from 0 to 60 (the scoring range of individual darts). The mean duration of all videos was 1.24 s ($SD = 0.49$ s). The software randomly chose 10 darts from the six experimental categories and presented them all in random order. Hence, every participant gave ratings for 60 stimuli presented in a different random order. We again, collected the same domain-specific dart's knowledge and experience variables as in Experiment 3.

Results

Study A The first missed dart was estimated on average as scoring 32.85 points ($SD = 6.18$), the second missed dart as scoring 31.67 points ($SD = 6.08$), and the third missed dart as scoring 32.19 points ($SD = 5.72$). Whereas, the first hit dart was estimated on average as scoring 34.15 points ($SD = 4.73$), the second hit dart as scoring 33.99 points ($SD = 5.28$), and the third hit dart as scoring 34.78 points ($SD = 7.42$). A 2 (hit vs. miss) $\times$ 3 (number of dart in the throw) repeated measures ANOVA on the mean estimated score of the respective darts revealed a significant main effect of hit (hit vs. miss) on the estimated score of the target players, $F(1, 56) = 12.316$, $p = .0009$, $\eta_p^2 = 0.180$. A sensitivity analysis

Fig. 5 Mean estimated scores of Experiment 4 as a function of hit versus miss and sequence of the darts on the pooled data of study a and the replication study b. Error bars represent standard errors

showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.20$. Neither the main effect of sequence order, $F(2, 112)=0.767$, $p=.467$, $\eta_p^2=0.014$, nor the interaction between hit and sequence order, $F(2, 112)=0.624$, $p=.538$, $\eta_p^2=0.011$, were significant.

Replication Study B The first missed dart was estimated on average as scoring 34.28 points ($SD=7.47$), the second missed dart as scoring 31.43 points ($SD=8.11$), and the third missed dart as scoring 33.54 points ($SD=7.59$). Whereas, first hit dart was estimated on average as scoring 35.84 points ($SD=7.06$), the second hit dart as scoring 34.48 points ($SD=9.51$), and the third hit dart as scoring 34.00 points ($SD=7.38$). A 2 (hit vs. miss) $\times$ 3 (number of dart in the throw) repeated measures ANOVA on the mean estimated score of the respective darts revealed a significant main effect of hit (hit vs. miss) on the estimated score of the target players, $F(1, 60)=7.144$, $p=.01$, $\eta_p^2=0.106$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.20$. This time there was also a main effect of sequence order, $F(2, 120)=3.462$, $p=.035$, $\eta_p^2=0.055$, which we do not have an explanation for. The interaction between hit and sequence order, $F(2, 120)=1.301$, $p=.276$, $\eta_p^2=0.021$, was again not significant.

Pooled Analyses A 2 (study_a vs. study_b) $\times$ 2 (hit vs. miss) $\times$ 3 (number of dart in the throw) mixed ANOVA on the mean estimated scores of the respective darts revealed a significant main effect of hit (hit vs. miss) on the estimated score of the target players, $F(1, 116)=18.779$, $p=.0001$, $\eta_p^2=0.139$. A sensitivity analysis showed that the test was sufficiently sensitive (power 0.80) to detect an effect of $f=0.10$. Further, there was a main effect of sequence order, $F(2, 232)=3.665$, $p=.027$, $\eta_p^2=0.055$, which we do not have an explanation for. No other main effects or interactions were significant, most notably the interaction between hit and sequence order, $F(2, 232)=0.926$, $p=.398$, $\eta_p^2=0.008$, was again not significant.

The descriptive statistics of the pooled analyses are summarized in Fig. 5.

Planned contrast between the individual darts revealed significant difference between all the hit and miss darts (first: $p = .041$; second: $p = .001$; third: $p = .036$).

Individual Difference Analyses As in Experiment 3, we decided to run additional individual difference analyses across both samples, as 118 participants would give reasonable power for the ANCOVA. None of the indicators of Darts experience significantly moderated the results (when entered as a covariate across both samples). The categorical variables gender and Darts level also did not significantly moderate the pattern of results when entered as a between-subject factor in the main analyses.

Discussion

The results of both versions of Experiment 4 confirm our hypothesis that the pre-performance nonverbal behavior of single darts hitting a triple led to overall higher performance ratings than the pre-performance behavior of darts not hitting a triple. The most important extension of Experiment 4 compared to the Experiments 1 and 3 was that it provided evidence that significant differences emerged between all three successful and unsuccessful darts within a throw (i.e., three darts thrown immediately after each other). Significant differences in the first darts of successful and unsuccessful series of three darts, suggest that participants also estimated higher scores based on pre-performance nonverbal behavior of a successful dart that was not immediately preceded by a successful prior dart. Hence, there does seem to be valid information in a professional darts player's pre-performance expressions that allows observers to find indications about how this player is likely to perform. Effect sizes for the main effects of hit vs. miss were medium-to-large by convention.

General Discussion

The central aim of this series of studies was to test the hypothesis that pre-performance nonverbal behavior of professional Darts players gives valid information to observers about tendencies concerning their subsequent performance. Results across four experiments, that all successfully replicated in direct replication attempts, supported this hypothesis. Results of Experiments 1–3 showed statistically significant linear trends between the score categories (poor performance, medium-to-good performance, good performance, and perfect performance) and mean performance ratings. However, closer examination of the contrasts between the individual performance categories suggest that participants could not distinguish equally well between all four performance categories, but could reliably distinguish between poor-to-medium performance and good-to-perfect performance. The results of Experiment 4 additionally show that the pre-performance behavior of one individual dart is sufficient to provide observers with valid cues about subsequent performance tendencies of professional darts players (importantly also in the first dart of a series of three darts that is not immediately preceded by a prior successful or unsuccessful dart). Hence, the present research is the first to show that highly skilled individuals seem to show nonverbal cues about their inner states that observers can pick up to draw accurate inferences about how this individual is likely to perform (in tendency).

The central finding of the series of experiments extends the thin-slices literature as it suggests that a person's momentary nonverbal behavior provides valid information to an

observer about how another person is likely to behave/perform in the next moment. To date, this has only been shown at larger times-scales (minutes, hours, months, years) in prior empirical research (e.g., Ambady et al. 2006; Curhan and Pentland 2007). As social interactions amongst humans are highly complex, it has been suggested that people have become especially attuned to nonverbal cues (e.g., facial expressions) of other people in order to facilitate social interactions (McArthur and Baron 1983; Zebrowitz and Collins 1997). In line with this theorizing, we consider it important to show that a person's nonverbal behavior has the potential to operate at very short time scales and inform other people about the next action of a person.

The findings of Experiment 3 suggest that facial information is sufficient to inform observers about performance tendencies of the darts players. Although we do not know the exact cues that observers used to inform their ratings, we speculated that valid information might be derived from the eyes of the actors given the large body of research showing contingencies between an athletes' gaze behavior and performance in aiming tasks (i.e., *Quiet Eye*; Rienhoff et al. 2016; Vickers 2007 for reviews). Based on the concept of nonverbal response system coherence (e.g., Matsumoto and Hwang 2013), which suggests that a person's nonverbal response to internal and external patterns of activation lead to an orchestrated response across different nonverbal communication channels (e.g., eyes, facial musculature, posture, etc.), it seems likely that, for example, a relaxed state (that has been linked to successful perceptual-motor performance; e.g. Weinberg and Gould 1999 for a review) of an athlete showed in various nonverbal channels and informed observer ratings. In this respect, future research would benefit from identifying the precise nonverbal behaviors that darts players display before performing successfully in comparison to less successfully. Therefore, future research might use existing coding schemes from other domains, like the Facial Action Coding System (Ekman and Friesen 1978) or the Body Action and Posture Coding System (Dael et al. 2012), to identify the (facial) movements and behaviors associated with successful sports performance.

The present data might be interpreted as in line with the behavioral ecology theory of facial expressions (Fridlund 1994; Crivelli and Fridlund 2018) because the nonverbal behavior in the stimulus material informed observers about how the expresser was likely to act in the immediate future. This is a basic tenet of the behavioral ecology theory of facial expressions (Fridlund 1994; Crivelli and Fridlund 2018). However, the present research was not set up as a focal test of the theory or can be regarded as evidence against rival theoretical accounts (e.g., basic emotion view; Darwin 1872; Ekman 1992; Keltner et al. 2019) as it did not attempt to test these theories against each other.

We consider it important to mention that our participants had little experience regarding darts. When we conducted analyses to find out if domain-specific experience (either playing or watching) darts moderated our main finding (i.e., the linear trend of performance ratings across performance categories) we did not find any evidence indicating this in Experiment 3 and 4. This finding is in line with previous research (Furley and Schweizer 2014) showing that domain-specific experience does not influence accuracy in drawing inferences from thin-slices of nonverbal behavior in sports.

The present research approach has some notable strengths and weaknesses. We consider it a strength that our stimulus materials were derived from actual sports competitions, instead of being artificially created, which is often the case in the field of nonverbal behavior. Thus, external validity should be rather high, and results are likely to transfer to similar field settings. Furthermore, the following steps were taken to maximize power of the present research. First, we made sure that sample sizes were adequate by computing a priori power sample sizes that would have sufficient power (0.95) to obtain

small-to-medium effects ($f=0.2$) following the recommendations of Faul et al. (2007). Second, all studies employed within-subject designs as a means of enhancing power (Open Science Collaboration 2017). Third, our dependent measure (estimated points scored) directly measured the variable of interest (how many points were scored) on the same scale. Fourth, in planning this research we addressed the problem of stimulus sampling (Wells and Windschitl 1999) by randomly selecting two different stimulus sets from the entire 2017 Professional Darts Corporation (PDC) World Darts Championship (12,056 throws). Finally, different participants were randomly shown different subsets of the stimulus material. This reduces the likelihood that results are dependent on one particular set of stimuli or sequence of stimuli. In addition, the present series of experiments are well-aligned with the increasing calls for replication in the psychological literature (Open Science Collaboration 2015; Simons 2014).

The main limitation of the present study is that it is unclear which cues perceivers used in the experiments, only that these cues are most likely in the faces of the players. Further, we chose to use stimuli from only darts and no other performance situations and, hence do not know if the present results only apply to professional darts. Finally, we have to acknowledge that the majority of research participants in the present series of studies were mostly W.E.I.R.D (western, educated, and from an industrialized, rich, and democratic country) from a worldwide perspective (Henrich et al. 2010). While we did not limit our data collection to college students (e.g., by the online studies in Experiment 1 and 2), the majority of participants were college students from the German Sport University Cologne. In this respect, it is worth mentioning that none of the demographic variables (gender, age, domain-specific experience) collected seemed to influence the general effect, we do consider the lack of diversity a limitation of the present study.

In conclusion, humans are constantly displaying nonverbal cues—whether they want to or not—that are associated with certain internal states and thereby inform other people on how they are currently feeling and likely to behave in the future (also in the very next moment). This basic theorizing seems to also apply to the pre-performance nonverbal behavior of professional darts players as observers can accurately infer performance tendencies from short recordings of facial expressions sampled immediately before the performance.

Research Disclosure Statements

The authors declare that (1) (a) the total number of excluded observations and (b) the reasons for making these exclusions have been reported in this manuscript; (2) that all independent variables or manipulations, whether successful or failed, have been reported in the manuscript; (3) that all dependent variables or measures that were analyzed for this article's target research question have been reported in in the manuscript/online supplements; (4) that (a) how sample size was determined and (b) our data-collection stopping rule have been reported in online supplements to this manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability All data will be made available at reasonable request to the corresponding author.

Compliance with Ethical Standards

Conflict of interest The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 32, pp. 201–271). San Diego: Academic Press.
- Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments. *Journal of Personality and Social Psychology*, 83, 947–961. <https://doi.org/10.1037//0022-3514.83.4.947>.
- Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, 16, 4–13. https://doi.org/10.1207/s15327663jcp1601_2.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 236–274. <https://doi.org/10.1037//0033-2909.111.2.256>.
- Argyle, M. (1972). *The psychology of interpersonal behaviour*. Harmondsworth: Penguin.
- Argyle, M. (1990). *Bodily communication*. London: Routledge.
- Baron-Cohen, S., Riviere, A., Fukushima, M., French, D., Hadwin, J., Cross, P., et al. (1996). Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition*, 3, 39–60. <https://doi.org/10.1080/713756728>.
- Bogdan, R. J. (1997). *Interpreting minds*. Cambridge: MIT Press.
- Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, 86, 599–614. <https://doi.org/10.1037/0022-3514.86.4.599>.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Altmeld, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2, 637–644. <https://doi.org/10.1038/s41562-018-0399-z>.
- Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41, 1054–1072. <https://doi.org/10.1016/j.jrp.2007.01.004>.
- Cozolino, L. (2006). *The neuroscience of human relationships: Attachment and the developing social brain*. New York: WW Norton & Co.
- Crivelli, C., & Fridlund, A. J. (2018). Facial displays are tools for social influence. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2018.02.006>.
- Crivelli, C., & Fridlund, A. J. (2019). Inside-out: From basic emotions theory to the behavioral ecology view. *Journal of Nonverbal Behavior*, 43, 161–194. <https://doi.org/10.1007/s10919-019-00294-2>.
- Curhan, J., & Pentland, A. (2007). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first five minutes. *Journal of Applied Psychology*, 92, 802–811. <https://doi.org/10.1037/0021-9010.92.3.8027>.
- Dael, N., Mortillaro, M., & Scherer, K. R. (2012). The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior*, 36, 97–121. <https://doi.org/10.1007/s10919-012-0130-0>.
- Darwin, C. (1872). *The expression of emotions in man and animals*. London: Murray.

- DePaulo, B. M. (1992). NVB and self-presentation. *Psychological Bulletin*, 111, 203–243. <https://doi.org/10.1037/0033-2909.111.2.203>.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6, 169–200. <https://doi.org/10.1080/02699939208411068>.
- Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11, 31–34. <https://doi.org/10.1177/1745691615596992>.
- Ekman, P., & Friesen, W. (1978). *Facial action coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. <https://doi.org/10.3758/BF03193146>.
- Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, 6, 163–171. <https://doi.org/10.1177/1745691611400237>.
- Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego: Academic.
- Furley, P. (2019). What modern sports competitions can tell us about human nature. *Perspectives on Psychological Science*, 14, 138–155. <https://doi.org/10.1177/1745691618794912>.
- Furley, P., & Schweizer, G. (2014). The expression of victory and loss: Estimating who's leading or trailing from nonverbal cues in sports. *Journal of Nonverbal Behavior*, 38, 13–29. <https://doi.org/10.1007/s10919-013-0168-7>.
- Furley, P., & Schweizer, G. (2016a). In a flash: Thin slice judgment accuracy of leading and trailing in sports. *Journal of Nonverbal Behavior*, 40, 83–100. <https://doi.org/10.1007/s10919-015-0225-5>.
- Furley, P., & Schweizer, G. (2016b). Nonverbal communication of confidence in soccer referees: An experimental test of Darwin's leakage hypothesis. *Journal of Sport and Exercise Psychology*, 38, 590–597. <https://doi.org/10.1123/jsep.2016-0192>.
- Furley, P., & Schweizer, G. (2020). Body language in sport. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (4th ed., pp. 1201–1219). Hoboken: Wiley.
- Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Schmid Mast, M., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality*, 42, 1476–1489. <https://doi.org/10.1016/j.jrp.2008.06.013>.
- Hall, J. A., Schmid Mast, M., & West, T. V. (Eds.). (2016). *The social psychology of perceiving others accurately*. Cambridge: Cambridge University Press.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83. <https://doi.org/10.1017/S0140525X0999152X>.
- Keltner, D., Tracy, J. L., Sauter, D., & Cowen, A. (2019). What basic emotion theory really says for the twenty-first century study of emotion. *Journal of Nonverbal Behavior*, 43, 195–201. <https://doi.org/10.1007/s10919-019-00298-y>.
- Kruglanski, A. W. (1989). The psychology of being “right”. The problem of accuracy in social perception and cognition. *Psychological Bulletin*, 106, 394–409. <https://doi.org/10.1037/0033-2909.106.3.395>.
- Matsumoto, D., Frank, M., & Hwang, H. S. (2013). *Nonverbal communication: Science and applications*. Los Angeles: Sage.
- Matsumoto, D., & Hwang, H. S. (2013). Facial expressions. In D. Matsumoto, M. Frank & S. H. Hwang, H. S. (Eds.), *Nonverbal communication: Science and applications* (pp. 15–52). Los Angeles: Sage.
- McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, 90, 215–238. <https://doi.org/10.1037/0033-295X.90.3.215>.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <https://doi.org/10.1126/science.aac4716>.
- Open Science Collaboration. (2017). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York: Wiley. <https://doi.org/10.1002/9781119095910.ch1>.
- Peirce, J. W., & MacAskill, M. R. (2018). *Building experiments in PsychoPy*. London: Sage.
- Rienhoff, R., Tirp, J., Strauß, B., Baker, J., & Schorer, J. (2016). The ‘quiet eye’ and motor performance: A systematic review based on Newell's constraints-led model. *Sports Medicine*, 46, 589–603. <https://doi.org/10.1080/17461391.2012.683815>.
- Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science*, 19, 109–111. <https://doi.org/10.1111/j.1467-9280.2008.02054.x>.
- Schmid Mast, M., & Hall, J. A. (2018). The Impact of interpersonal accuracy on behavioral outcomes. *Current Directions in Psychological Science*, 27(5), 309–314. <https://doi.org/10.1177/0963721418758437>.

- Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, 23, 114–122. <https://doi.org/10.1016/j.psychsport.2015.11.005>.
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9, 76–80. <https://doi.org/10.1177/1745691613514755>.
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308, 1623–1626. <https://doi.org/10.1126/science.1110589>.
- Unkelbach, C., & Memmert, D. (2010). Crowd noise as a cue in referee decisions contributes to the home advantage. *Journal of Sport and Exercise Psychology*, 32, 483–498.
- Vickers, J. N. (2007). *Perception, cognition, and decision training: The quiet eye in action*. Champaign: Human Kinetics.
- Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). *Pragmatics of human communication*. New York: W. W. Norton.
- Weinberg, R. S., & Gould, D. (1999). *Foundations of sport and exercise psychology* (2nd ed.). Champaign: Human Kinetics.
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychological Bulletin*, 25, 1115–1125. <https://doi.org/10.1177/01461672992512005>.
- Willis, J., & Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological Science*, 17, 592–598. <https://doi.org/10.1111/j.1467-9280.2006.01750.x>.
- Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review*, 1, 204–223. https://doi.org/10.1207/s15327957pspr0103_2.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.