Aalto University

Chang-Arana, Alvaro M.; Piispanen, Matias; Himberg, Tommi; Surma-aho, Antti; Alho, Jussi; Sams, Mikko; Holtta-Otto, Katja

Empathic accuracy in design

Design Science

Empathic accuracy in design: Exploring design outcomes through empathic performance and physiology

Álvaro M. Chang-Arana[1], Matias Piispanen[2], Tommi Himberg[2],
Antti Surma-aho[1], Jussi Alho[2], Mikko Sams[2,3] and Katja Hölttä-Otto[1]

1 *Department of Mechanical Engineering, Aalto University, 02150, Betonimiehenkuja 5 C, Espoo P.O. Box 17700, FI-00076 AALTO, Finland*

2 *Department of Neuroscience and Biomedical Engineering, Aalto University, 02150, Betonimiehenkuja 5 C, Espoo P.O. Box 17700, FI-00076 AALTO, Finland*

3 *Department of Computer Science, Aalto University, 02150, Betonimiehenkuja 5 C, Espoo P.O. Box 17700, FI-00076 AALTO, Finland*

the **Design Society**
a worldwide community

CAMBRIDGE
UNIVERSITY PRESS

Abstract

Empathic design highlights the relevance of understanding users and their circumstances in order to obtain good design outcomes. However, theory-based quantitative methods, which can be used to test user understanding, are hard to find in the design science literature. Here, we introduce a validated method used in social psychological research – the empathic accuracy method – into design to explore how well two designers perform in a design task and whether the designers' empathic accuracy performance and the physiological synchrony between the two designers and a group of users can predict the designers' success in two design tasks. The designers could correctly identify approximately 50% of the users' reported mental content. We did not find a significant correlation between the designers' empathic accuracy and their (1) performance in design tasks and (2) physiological synchrony with users. Nevertheless, the empathic accuracy method is promising in its attempts to quantify the effect of empathy in design.

Key words: Empathic design, empathy, empathic accuracy, user understanding, EMG

1. Background

1.1. Empathy in design and engineering

The use of the term *empathy* has steadily increased over the past two decades in academic journals dealing with the business world (Köppen & Meinel 2015). This term is widely used in design approaches such as *human-centred design* or *design thinking*, both of which have been associated with successful projects or businesses (Brown 2009; Kramer, Agogino & Roschuni 2016). However, there is no widely accepted and consistently used definition of *empathy* in design. Empathy is defined in multiple ways: as a mindset, as a way of understanding others, as a method or as behaviour.

Extensive literature reviews (Kouprie & Sleeswijk Visser 2009; Strobel *et al.* 2013; Walther, Miller & Sochacka 2017), borrowing definitions from

psychology (Wong *et al.* 2016; Surma-aho, Björklund & Hölttä-Otto 2018) and based on interviews with designers (Strobel *et al.* 2013; Hess, Strobel & Pan 2016) and observing designers (Hess & Fila 2016) suggest that empathy is commonly equated with some type of *comprehensive user understanding*. For instance, empathy in design has been associated with user-understanding methods like immersing oneself in the dreams of a future user (Battarbee *et al.* 2002), imposing extreme user-like features on designers (Vaughan, Seepersad & Crawford 2014; Pang & Seepersad 2016) or on non-extreme users (Lin & Seepersad 2007), understanding users through a combination of survey and sensor data (Ghosh *et al.* 2017), and projecting into a user's life through using one's imagination (Koskinen & Battarbee 2003). Some studies define designer empathy as an outcome of user interaction – an increased ability to understand users and solve their issues (Raviselvam *et al.* 2017; Raviselvam *et al.* 2018). However, it is not clear when and how user understanding can be considered empathic. Some studies have attempted to clarify this situation by adopting more rigorous definitions of designer empathy, typically based on psychology research. One notable conceptualisation of designer empathy was developed by Kouprie and Sleeswijk Visser (2009). They depicted a stepwise structure for designers to develop and use empathy with end users that involved the designer putting herself or himself in situations typical for the end user and doing tasks as if they were the user, eliciting information directly from users through various types of interaction and combining these two sources of information to achieve comprehensive and empathic understanding (Kouprie & Sleeswijk Visser 2009).

Several other aspects of empathy inherent to design have been identified. Experienced designers value empathy more than their younger colleagues (Hess *et al.* 2017). Designers should empathise with both their peers and the end users (Strobel *et al.* 2013; Köppen & Meinel 2015). Designers should alternate between empathic thinking and analytical thinking (Walther *et al.* 2017). It has also been suggested that empathy for users is not only important when designers are gathering user information but also during other activities such as requirement definition and concept generation (Hess & Fila 2016). However, both instructions to active designers (IDEO 2015) and preliminary case studies (Smeenk, Tomico & Van Turnhout 2016) indicate that to develop successful products, designers must use their own insight in combination with comprehensive user understanding. Ultimately, even when what *empathy* comprises and how it is created are not well defined, all depictions of empathy in design agree on the aim of achieving an accurate, comprehensive understanding of the user and using this understanding to make future design decisions.

In psychology, empathy is not fully understood. However, it is usually conceptualised as a bidimensional construct including cognitive empathy and affective empathy (Shamay-Tsoory 2011). Cognitive empathy involves top-down processes that allow an individual to imagine and cognitively share what someone else could be thinking or feeling. In design, cognitive empathy is usually understood as perspective taking (Koskinen & Battarbee 2003; Postma *et al.* 2012; Köppen & Meinel 2015). Affective empathy involves bottom-up processes that allow an individual to recognise someone else's emotions and even share similar or equal emotional states. It includes several mechanisms such as emotional contagion (Preston & de Waal 2002), sharing the experience of pain or distress with others (Singer *et al.* 2004; Jackson, Meltzoff & Decety 2005), reacting to

someone else's facial expressions (Carr *et al.* 2003), empathic concern (Light *et al.* 2015) etc. Furthermore, autism and psychopathy research suggests that in some clinical cases, individuals may only have the capacity for one form of empathy (Baron-Cohen & Wheelwright 2004; Bird & Viding 2014; Ellis *et al.* 2017; Moreira, Azeredo & Barbosa 2019). But this division is not clear since the two components of empathy interact more than previously thought (Cuff *et al.* 2016). For instance, empathic concern has been measured using questionnaires such as the Interpersonal Reactivity Index (IRI, Davis 1980) and thus asking for the top-down reasoning for how a person generally feels when perceiving someone else in distress or pain. However, empathic concern (and other mechanisms) can also be measured through bottom-up procedures such as physiological synchrony. This approach has been used in fields such as psychotherapy (Kleinbub 2017) and dyadic interactions between parent–child and couples (Palumbro *et al.* 2017) in order to measure different outcome variables such as the patient rating of a therapist's empathy (Marci *et al.* 2007) and the occurrence of child behavioural problems (Lunkenheimer *et al.* 2015) and marital conflict (Gates *et al.* 2015).

In addition to problems brought about by the ambiguity in the definition of empathy, another key limitation in current empathy research in design is the lack of quantitative studies connecting empathy to design outcomes. Quantitative studies could be used to create predictive models of how empathy – be it defined as a mindset, understanding, method or behaviour – influences design outcomes. Existing quantitative research on empathy in design has used validated self-report measures from psychology to show that design students learn empathy in project classes (Surma-aho *et al.* 2018) and that engineering students typically have lower dispositional empathy than students of psychology and social work (Rasoal, Danielsson & Jungert 2012). Another notable example is the Empathy and Care Questionnaire, which is used to assess practitioners' self-reported perceptions of empathy (Hess *et al.* 2017). However, just a few quantitative studies have been carried out on empathy in design, and no research has truly tested whether empathy translates into improved design outcomes such as correct needs understanding, better ideas, user satisfaction, product usability or perceived effectiveness.

1.2. From empathy in design to empathic accuracy in design

Empathy in design is targeted towards a specific user group in a specific context. For instance, designers working with a group of musicians will try to understand their pains and joys, likes and dislikes about their instruments or playing music. This understanding entails careful observation, interviews aimed at uncovering different nuances of their context and other methods that can be used to inform decision-making. Therefore, studying empathy in design is challenging in the sense of establishing general rules for good approaches given its context-specific nature.

Most research providing important information about the role of empathy in design is qualitative (e.g. Kouprie & Sleeswijk Visser 2009; Kankainen *et al.* 2012; Smeenk *et al.* 2017). While the qualitative approach allows us to delve into the specific context and understand the experience of the agents involved within, it does not allow us to make quantitative predictions. Therefore, qualitative approaches need to be complemented by quantitative ones that allow us to predict, explain and control the role of empathy in the design process.

The empathic accuracy method is a performance-based method for measuring the degree of understanding between two or more people interacting in a specific context in real time. It provides a quantitative measurement of the understanding of another person without self-rating empathic skills. There are three versions of the paradigm, all of which require video recording a conversation between a dyad (e.g. a user and a designer): the dyadic interaction paradigm, the standard stimulus paradigm and the shared physiology paradigm. The first two estimates the degree of similarity between lists of mental contents provided by either or both member of a dyad or from external perceivers of the interacting dyad. The higher the similarity of reported mental contents, the higher the understanding between the members of the dyad or between perceivers and members of the dyad. The third paradigm make use of physiological synchrony instead of reported mental contents to estimate the understanding between members of a dyad or perceivers. This paradigm equates higher physiological synchrony with higher accuracy when inferring someone else's feelings. Given the task of each paradigm, they may lie closer to the cognitive or affective component of empathy. Broadly, we can locate the dyadic interaction paradigm and the standard stimulus paradigm under the cognitive empathy component and the shared physiology paradigm under the affective empathy component.

1.2.1. The dyadic interaction paradigm

In the dyadic interaction paradigm (Ickes *et al.* 1990), the members of a dyad are separately asked to rewatch their videoed interaction. One of the participants is asked to pause the recording every time they remember having had a specific thought or feeling during the interaction and to write down this thought or feeling. The second member of the dyad then watches the same video, but now it pauses at the same time points where the first participant paused it to report a specific thought or feeling. The second participant must write down what they think the first person was thinking or feeling. The two lists are compared by a group of independent participants who rate how similar the items on the two lists are. The higher the similarity, the higher the *empathic accuracy* of the second participant.

1.2.2. The standard stimulus paradigm

In the standard stimulus paradigm (Marangoni *et al.* 1995), the videoed dyadic interaction is used as a standard stimulus from which a group of perceivers infer the thoughts and feelings of either one of or both dyad members. The perceivers do not have direct contact with either one of members of the videoed dyad.

The advantage of the dyadic interaction paradigm and the standard stimulus paradigm is that they allow us to directly compare what a user thinks or feels with what a designer thinks the user is thinking or feeling. Importantly, it also allows the study of whether the measured accuracy is similar to a designer's self-rated accuracy in regard to identifying a user's mental contents. Previous studies have shown that people tend to have a low degree of empathic accuracy (Stueber 2018). For instance, when inferring another person's thoughts or feelings, an approximate accuracy of 20% was achieved between strangers and about 30% between people who had known each other for at least one year (Ickes & Hodges 2013). Obviously, people are rather bad at inferring what someone is thinking or feeling when the topic of the discussion is open.

Because participants are instructed to infer what someone else might be thinking or feeling, these tasks measure cognitive empathy. That is, they measure imagining someone else's thoughts and feelings in a given circumstance or seeing the world from someone else's psychological perspective (Shamay-Tsoory 2011; Zaki & Ochsner 2012). In this case, seeing the world from someone else's perspective is operationalised as the degree of similarity between the actual mental contents and inferred contents, a similar concept to Davis' perspective-taking factor on his IRI (1980).

1.2.3. *The shared physiology paradigm*

The other version of the empathic accuracy method also records an interacting dyad, with the addition of monitoring physiological responses (such as heart rate, skin conductance and facial muscle activity) to capture affective empathy (Levenson & Gottman 1983; Levenson & Ruef 1992). Modern versions of this paradigm have incorporated brain imaging as an additional measure of affective empathy (Zaki *et al.* 2009). In essence, the paradigm measures how accurately a participant identifies the ongoing feelings of someone else; the synchrony of physiological responses is used to estimate the similarity of felt emotions (Levenson & Ruef 1992).

Interest in physiological synchronisation has increased in recent years in both psychology and neuroscience studies (Kreibig 2010; Quintana & Heathers 2014; Massaro & Pecchia 2019). Studies on social interactions show that physiological, behavioural and emotional reactions tend to be shared or synchronised during interaction. Synchronisation has been observed in situations such as recognising the emotions of a person from another culture (Soto & Levenson 2009), the interaction of married couples (Levenson & Gottman 1983) or simply sharing the same space while watching emotional movies (Golland, Arzouan & Levit-Binnun 2015).

2. The current study

This study aims at addressing the shortage of quantitative studies connecting empathy to user understanding in a specific design context and testing whether empathy is relevant for design. We combined elements from all of the above-mentioned paradigms in order to study if empathic accuracy plays a role in an early-phase design and ideation task.

Within this context, our aim was to measure empathic accuracy as a quantitative indicator for a designer's empathic capability. We analysed the interaction between two professional designers and five musicians. We formulated the following research questions:

(1) How accurately can the designers understand the group of musicians?
(2) Does the designers' accuracy in regard to the musicians' mental contents and emotions positively correlate with design outcomes?
(3) Does the similarity of the facial emotional expressions of the designers and musicians correlate with the designer's empathic accuracy?

3. Method

3.1. Participants

Two designers were recruited. The interviewing designer (Designer 1) had 13 years of experience, including six years of design education (gaining a bachelor's degree and a Master of Science degree), four years of human-centred design work and three years of design research. He also received weekly teaching from a piano instructor for 12 years when he was young. Although he had no professional training on the instrument, his previous experience in music was assumed to be an important requirement for understanding the musicians. In addition, the first author of this study assisted the designer during the planning phases of the interviews. He has played piano for 15 years and has a Master of Arts degree in music psychology, which aided in formulating relevant interview questions.

The second designer (Designer 2), a co-author of this study, had 5.5 years of experience, including three years of design education (MS in product development) and 2.5 years of design research. He did not have musical education except that gained in regular primary school. He was asked to watch the interviews and perform the dyadic interaction paradigm task for two reasons. First, to control for the effects that Designer 1's design experience and musical background could have on his performance. Second, to test whether indirect contact with the users would translate into considerably different empathic accuracy scores compared with those obtained by Designer 1.

Five professional musicians (three females: two clarinettists, two saxophonists and one oboist; mean age $= 23.60$, $SD = 1.52$) with a mean playing-time experience of 15 years ($SD = 1.41$) participated. The musicians were recruited using their musical institution's mailing list. The musicians belonged to four different nationalities and only one had English as her mother tongue. The rest had at least B1 level English according to the Common European Framework of Reference for Languages, as demanded by their music institution.

3.2. Design brief

The designers' task was to understand and ideate accessories to improve the musicians' experiences with their instruments. The musicians involved in the study were professional woodwind players, many of whom experience similar challenges related to their instruments, most importantly those associated with the use of reeds. Reeds are small strips of wood or plastic that vibrate with air pressure and influence the airflow into the instrument. They affect the production of tone and the expressive and technical range of the musician (Thompson 1979; Ledet 1981; Almeida *et al.* 2013). Besides these music-related features, reeds present additional challenges such as their limited lifespan, the personal preference of each musician, the high cost of purchasing them or manufacturing them from scratch and the considerable amount of time reed making takes (Ledet 1981). The problems around reeds and their potential impact on the performance and well-being of musicians (Nagel 2010; Kenny 2011) are an important challenge for design. In addition to reeds, the designers were given the freedom to focus on other accessories that the musicians might need such as solutions for transporting and storing their instruments or cleaning equipment. This design brief and the associated tasks, while not spanning the entire design process, provide a realistic
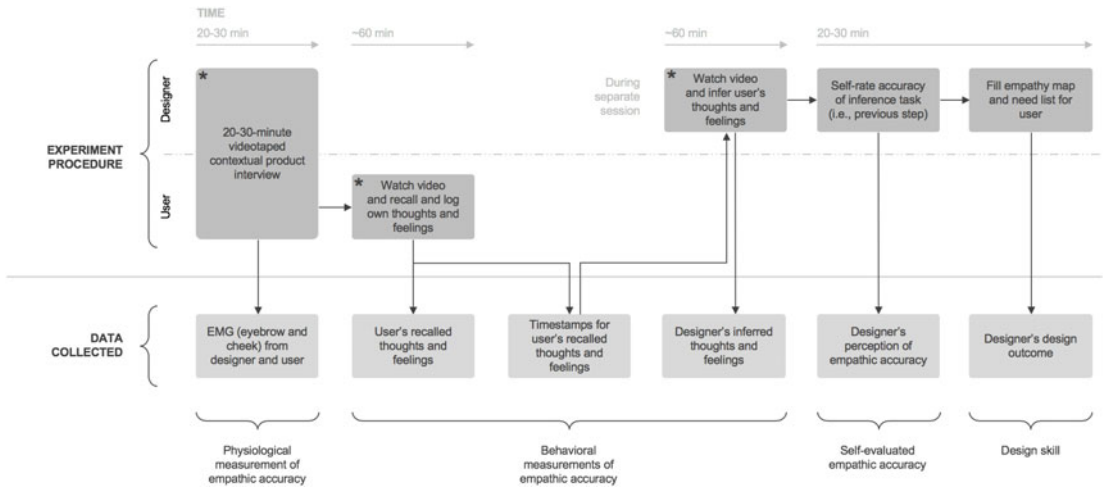
Figure 1. An overview of the study procedure.

starting point and a set of initial actions taken by design practitioners in various open-ended projects.

3.3. Tasks and procedures

Before describing our methods in detail, we describe a simplified version for illustration. From a videoed interview between a designer and a user wearing physiological electrodes, two lists of mental contents are obtained: a list of remembered mental contents from the user and a list of inferred mental contents from the designer. These lists are rated on their content similarity by external raters, thus assigning an 'empathic accuracy' to a designer. Then, the designer completes two design tasks that are rated by the interviewed user. Behavioural and physiological outcomes are correlated with design outcomes in order to test whether higher empathic accuracy and physiological synchronisation correlate with higher performance in design outcomes. An overview of the approach is shown in Figure 1.

3.3.1. Interview

Designer 1, together with the first author, developed guidelines for a 20–30-minute semi-structured interview (see Appendix 1). Designer 1 was in charge of conducting the interview given his extensive design and needs-finding experience, and the first author did not participate in it. During the interview, the musicians manipulated their instruments for demonstration purposes. This included, for example, setting up the instrument for playing and demonstrating cleaning the instrument. This was done to mimic a more contextual interview. Both Designer 1 and the musician, and later Designer 2, wore the same set of physiological sensors to record an electrocardiogram (ECG), facial electromyography (EMG) and galvanic skin response (GSR). Designer 2 was presented with the design brief and encouraged to place himself in the position of the interviewer and watch the interaction from a design perspective. We focus here on the EMG signals from the designers' and musicians' eyebrow muscles (*corrugator supercilii* muscles) and

7/34

cheek muscles (*zygomaticus major* muscles). The activity of these muscles serves to provide indices for frowning and smiling (proxies for negative and positive emotional valence, respectively).

Before starting each major phase of the study, participants filled in the Positive and Negative Affect Schedule (PANAS; Watson, Clark & Tellegen 1988). It was used to gauge the participants' emotional states before the interview and before completing two empathic accuracy tasks (here we only report Ickes' dyadic interaction paradigm). Because a single session with one musician lasted approximately four hours, we needed to control their mood in order that it was as constant as possible and would not be a confounding factor in their task performance. Designer 1 only spend 30 min per musician at this stage of the study. Thus, we assumed that fatigue would not have a noticeable detrimental effect on his performance and did not control for his mood changes. Before starting the interview, the participants were reminded about the topic of the interview and its approximate duration. Then they were instructed to be silent with their eyes closed for three minutes in order to stabilise their physiological signals. During the interview, the physiological signals from both members of the dyad were continuously recorded.

The interview with User 3 (U3) had to be restarted after the first five minutes due to an unexpected problem with the recording equipment. After solving the problem, the interview was resumed by summarising the prior discussion's content. The interview lasted for a total of 15 minutes. Data from U3 was otherwise collected and included in the analysis in a similar way to other users.

3.3.2. *Logging in remembered mental contents: The musicians' phase*

Before starting, the participants filled in PANAS once again. Following Ickes' validated protocol (2001), the musicians were asked to pause the video every time they remembered having a specific thought or feeling. They had to write down the thoughts and feelings they remembered instead of new thoughts or feelings that they might have while rewatching the video. The participants were presented with a practice trial and instructed in how to use the standard thought-or-feeling sheet. They were asked to write down the timing of where they paused the video, write down the content and choose whether it was a positive, neutral or negative thought or feeling. Although the emotional-valence choice does not allow identifying the exact emotion (i.e., choosing a negative emotion for a specific entry could be due to either anger or frustration), each valence choice is paired with specific content, detailed by the participant. Thus, it is possible to infer what the actual emotional experience was using the entry's content. The experimenter ensured participants fully understood the task before instructing them to begin as soon as the video started. Responses were registered using a digital standard response sheet based on Ickes' design (2001). Instructions were presented in printed form and answers were registered on the 'inferred thoughts or feelings' response sheet (see Appendixes 2A and 2C for examples of the instructions and response sheets).

3.3.3. *The dyadic interaction paradigm: Designer 1's phase*

Before starting this phase, Designer 1 was only aware of the objective of the interview, that it will be video recorded and that he will rewatch it while performing an unspecified task. After obtaining the list of thoughts and feelings from the musicians, Designer 1 was invited to rewatch the five interviews

approximately one month after the first interview and three days after the last interview. Before starting the task, Designer 1 filled in the PANAS. He was then instructed to infer as accurately as possible what a particular musician was thinking or feeling when she reported her thoughts and feelings as well as inferring what the emotional valence of that specific entry would be. At this point, Designer 1 was aware that every time the video was paused it was annotated by each musician since knowing this was a crucial requirement for this phase.

3.3.4. The standard stimulus paradigm: Designer 2's phase

Designer 2 completed the same task described above approximately five months after the last interview and was not aware of the content reported by the musicians or Designer 1 at the time of completing it. Distinctively, Designer 2 did not have direct contact with any of the musicians. Instructions were presented in printed form and answers were registered on the 'inferred thoughts or feelings' response sheet (see Appendixes 2B and 2D for examples of the instructions and response sheets).

3.3.5. Assessing the similarity of contents

Fourteen native speakers of English with completed undergraduate education or a higher level of education were recruited to rate the similarity of the content: the remembered thoughts and feelings, and the inferred thoughts and feelings of both designers (eight for Designer 1 and six for Designer 2). Following Ickes' protocol (2001), the similarity of content was assessed using a three-point Likert scale, ranging from 0 to 2. Raters assigned a 0 if both lists had 'essentially different content', 1 if they had 'somehow similar, but not the same content' and 2 if it was 'essentially the same content'. Six examples (two for each possible rating) were presented along with the instructions in order to clarify the meaning of each value. The raters were presented with the five pairs of lists of mental content in a randomised order. Reliability analysis followed Ickes' procedure. Each rater was treated as a questionnaire item and every entry score as a questionnaire response. Cronbach's alpha was then calculated for each interview. Nunnally's reliability criterion of .70 (1967) was used to assess the reliability of the obtained scores. Instructions were presented in digital form and answers were registered likewise (see Appendix 2E for an example of the instructions and response sheets).

3.3.6. The designer's self-rated performance in regard to the dyadic interaction paradigm

After the dyadic interaction task, Designer 1 was asked to rate how well he thought he had completed the task on a single-item 10-point Likert scale. Since Designer 2 was aware of the results of Designer 1's self-rated performance, he did not self-rate his performance.

3.3.7. An empathy map and ideas for improvements: The designers' phase

The designers were asked to create an empathy map to summarise and synthesise the key insights they could identify after participating in and rewatching the interviews. The empathy map was a modified version of Both and Baggereor's map (no date). Although this design tool contains four quadrants (i.e., 'say', 'do', 'think' and 'feel'), only 'think' and 'feel' were used in this study in order to enable a similar

comparison between the empathic accuracy scores and this design outcome. The empathy map's thoughts and feelings differ from those of an empathic accuracy task. The empathy maps contained general judgements about what the users might be thinking or feeling and were completed after the designers watched the interviews. There was no specific mental content tied to specific time occurrences. The designers also generated ideas for new and/or improved accessories for the musicians. The designers listed their ideas in a text after completing the empathy map task. They were encouraged to complete both tasks as if they were part of a professional design project. Both tasks were used to crudely mimic what the next steps in a real design case might be: synthesising user understanding and generating initial ideas for further development. The designers took roughly 30 minutes per interview to complete this phase. Instructions were presented in printed form and answers were registered on a standard response sheet (see Appendix 3).

3.3.8. The empathy map and ideas for improvements: Rating the empathy map and ideas for improvements

Insights from the empathy maps and the lists of ideas for improvements suggested by both designers were sent back to the musicians for rating. This was to simulate the design-process step of coming back to the user to obtain direct feedback on the initial ideas. The musicians used a five-point Likert scale to rate how close every insight in the empathy map's thoughts/feelings was to their experience as users. Similarly, the musicians rated the relevance of the proposed ideas using a five-point Likert scale, based on what they discussed during the interview. After completing these tasks, the musicians were fully debriefed about the aims of the study.

3.4. Materials

3.4.1. Data Logger

EMG data was collected using the portable telemetry and 16-channel data logger Biomonitor ME6000. The system allows the collection of different types of data including EMG, GSR and ECG data. EMG electrodes were placed on the left *corrugator supercilii* muscle and the left *zygomaticus major* muscle of each dyad member.

3.4.2. FSenSync (Förger Analytics)

A free-access software package was used to synchronise the recordings and streaming of the measured data. The software allows real-time streaming, recording, making notes, synchronising sensor units and compensating for slight clock drifts that may occur while recording.

3.4.3. Video recording

Interviews were videoed using Android phones running a video recording application synchronised to the FSenSync software. Cellphone cameras were placed at approximately the same height as the interviewer's and the musicians' eye level. The aim was to capture, as closely as possible, a frontal vision of each member of the dyad. The participants were framed from their seat upwards to ensure their hands and faces were visible at all times.

3.5. Data processing

3.5.1. An aggregated index of empathic accuracy

An empathic accuracy score was calculated for the performance of both designers in each interview following Ickes' procedure (2001). First, an average accuracy score for each entry was calculated. Second, a total index score was calculated by adding the average score for all entries. Third, the total index score was divided by the total amount of entries on each dyad in order to 'yield an index of the proportion of accuracy points relative to the total number of accuracy points possible' (p. 232). Fourth, indices of the proportion of accuracy points were percentage scaled by dividing them by two and then multiplying them by 100.

3.5.2. Electromyography preprocessing

The EMG signal was bandpass filtered at 20–400 Hz. A fast Fourier transform with a 1 s Hanning window and 0.5 s overlap was applied to filtered data in order to calculate power spectral density estimates (van Reekum *et al.* 2010; Lapate *et al.* 2014; Golland *et al.* 2018). The estimates were averaged and z-transformed to take account of variations in amplitudes between subjects.

3.5.3. Rating emotional valence

The users' reported emotional valences and the designers estimates of them were compared. When they coincided, this was scored as 1.

3.5.4. Cross-correlation analysis of muscle activity

The maximum cross-correlation within a ± 5 s lag was calculated for every 10 s time event window to determine the similarity between the designer's and users' facial expressions during an event.

3.5.5. Correlation of EMG and empathic accuracy

To calculate whether physiological synchrony between the designers and the users was related to the former's empathic accuracy, a Pearson correlation coefficient was calculated between the reactions of the *zygomatic major* (the 'smile muscle') and the empathic accuracy score obtained for each of the events where a thought or feeling was reported (117 entries).

3.5.6. Interpretation of effect sizes

Effect sizes (*r* or *rho*) were interpreted according to Cohen's criterion (see Ellis 2010): a small effect $= .10$, a medium effect $= .30$, a large effect $= .50$.

4. Results

Our data consists of five video interviews of about 30 minutes. We used eight different channels to collect physiological data from our participants including GSR, EMG and ECG data (although here we only report the results from EMGs). Additionally, 117 remembered thoughts and feelings were reported by the musicians, as were the corresponding inferences from both designers. Both designers reported a total of 169 thoughts and feelings on empathy maps and 43 ideas for improvements. Their relevance was assessed by the musicians.

Table 1. The inter-rater reliability of the assessment of the similarity of content

| | Designer 1 | | Designer 2 | |
	Cronbach's α	SEM	Cronbach's α	SEM
User 1	.90	0.09	.88	0.09
User 2	.90	0.14	.92	0.15
User 3	.88	0.15	.87	0.17
User 4	.86	0.14	.75	0.12
User 5	.90	0.12	.87	0.12

Note: User 1, entries $= 45$; User 2, entries $= 18$; User 3, entries $= 15$; User 4, entries $= 17$; User 5, entries $= 22$. Designer 1 was rated by eight external raters; Designer 2 by six external raters. SEM $=$ Standard Error or Measurement.

4.1. Controlling for change in the emotional state of users

A Wilcoxon signed-rank test indicated that the mean ranked emotional states of musicians were not significantly different between the beginning of the interview and the beginning of Ickes' empathic accuracy task in either the positive mood subscale ($z = -1.83$, $p = .07$, $r = -.58$) or the negative mood subscale ($z = -0.37$, $p = .72$, $r = -.03$). Thus, the musicians felt similarly in both conditions, and it is less likely that their performance on the empathic accuracy task was affected by mood changes. It was important to control the musicians' mood state given that the duration of the experiment was approximately four hours. For the designers, the experiment was much shorter – approximately 1 h 30 min. We assumed that fatigue would not have a noticeable detrimental effect on the designers' performance and did not control for their mood changes. However, some other things could have affected the designers' performance. Thus, controlling for their mood would have been important.

4.2. How accurately can designers understand a group of musicians?

4.2.1. The inter-reliability of the scoring of the similarity of content

Table 1 summarises the inter-rater reliability of the external raters' ratings for the similarity of content between the users' remembered thoughts and feelings, and both designers' inferred thoughts and feelings. The reliability values were above Nunnally's criterion of .70.

4.2.2. The designers' empathic accuracy score

The designers' aggregated index of empathic accuracy, self-rated accuracy when performing the empathic accuracy task (just Designer 1) and the percentage of correctly identified user emotional valence are summarised in Table 2. We tested if the designers' empathic accuracy differed significantly by using three Mann–Whitney tests: (1) the designers received similar scores from the external raters: $U = 6403.00$, $N = 234$, $z = -0.856$, $p = .39$, $r = -.06$; (2) the designers had similar aggregated indices of empathic accuracy: $U = 6.00$, $N = 10$, $z = -1.36$, $p = .18$, $r = -.43$; and (3) the designers performed similarly when

Table 2. The overall designers' empathic accuracy scores

	Designer 1			Designer 2	
	Aggregated index of empathic accuracy (%)	Designer's reported self-rated accuracy (%)	Correct identification of users' emotional valence (%)	Aggregated index of empathic accuracy (%)	Correct identification of users' emotional valence (%)
User 1	45.42	90.00	42.22	42.22	40.00
User 2	50.35	80.00	55.56	55.09	50.00
User 3	48.75	60.00	40.00	49.44	20.00
User 4	44.49	80.00	41.18	55.88	35.29
User 5	45.17	80.00	50.00	53.41	40.91

Note: Self-efficacy ranged from 1 to 10, here rescaled to percentage for ease of comparison.

identifying the users' emotional valence: $U = 5.00$, $N = 10$, $z = -1.58$, $p = .12$, $r = -.50$.

4.2.3. Examples of remembered and inferred thoughts and feelings
The 117 entries obtained from the five interviews were assessed by naive raters with scores from 0 to 2, ranging from *totally different content* to *essentially the same content*. Here we present examples of high-, mid- and low-performance accuracy for both designers as well as the emotional valence remembered by the musicians and inferred by the designers.

4.2.4. The development of empathic accuracy over time
To test whether the designer's empathic accuracy developed over time, we performed a Wilcoxon signed-rank test to compare the empathic accuracy scores obtained during the first and last 10 minutes of all interviews. Interview time did not have an effect on either designer's empathic accuracy. Designer 1's empathic accuracy for the first 10 minutes ($n = 49$, $Mdn = 0.88$, $SD = .58$) and last 10 minutes ($n = 37$, $Mdn = 1.13$, $SD = .56$) did not increase as the interviews progressed over time ($N = 74$, $z = -.07$, $p = .95$, $r = -.01$), even when excluding the third user from the analysis (due to the shorter duration of the interview): $N = 71$, $z = -.21$, $p = .84$, $r = -.02$. Similarly, a second Wilcoxon signed-rank test was done to compare the empathic accuracy scores obtained during the first 10 minutes ($n = 49$, $Mdn = 1.00$, $SD = .59$) and last 10 minutes ($n = 37$, $Mdn = 1.00$, $SD = .61$) from all the interviews watched by Designer 2. Likewise, it did not show significant changes ($N = 74$, $z = -.17$, $p = .87$, $r = -.02$), even when excluding the third user from the analysis: $N = 71$, $z = -.46$, $p = .65$, $r = -.01$.

4.2.5. Design task scores
Table 4 summarises the scores obtained by the designers in the three design tasks. The scores given by each musician were transformed into a percentage for ease of interpretation.

13/34

Table 3. Examples of high-, mid- and low-empathic accuracy

	Musicians	Designers	Average EA
High-empathic accuracy	I was feeling unprofessional/ashamed because I do not use the 'proper' boxes to keep my reeds safe. (−)	D1: She was feeling ashamed about not taking as good care of the reeds as she knows she could. (−)	1.88
	I was feeling amused by the thought of carrying a giant case. (+)	D2: She was remembering someone struggle or complain about not being able to take their clarinet into pieces and having to carry around a big backpack for it. (0)	1.83
Mid-empathic accuracy	I was thinking that I try to adapt to the reed, but I am not good enough yet. (−)	D1: She was feeling humble about not wanting to say she is good at adapting. (+)	1
	I was thinking that the neck strap prevents me to bow towards the audience in an elegant way. It could fall while bending forward. (0)	D2: She was recalling the relief of taking the shoulder strap off and being free of the sax's weight. (+)	1
Low-empathic accuracy	I was a bit frustrated as I started to struggle with English a bit in my head. (−)	D1: She was thinking back of how she got from hardly playing to playing more and more. (0)	0
	I was feeling happy, I like the word 'wizardry'. (+)	D2: She was thinking about the bad performance she had last week and how the reed felt then. (0)	0

Note: D1 = Designer 1; D2 = Designer 2. Negative valence (−), neutral valence (0) and positive valence (+).

Table 4. The designers' performance in three design tasks

	Designer 1			Designer 2		
	Empathy map		Ideas for improvements (%)	Empathy map		Ideas for improvements (%)
	Thoughts (%)	Feelings (%)		Thoughts (%)	Feelings (%)	
User 1	87.20	72.00	80.00	77.14	68.57	60.00
User 2	96.60	86.60	76.60	89.09	97.14	96.00
User 3	100.00	95.60	90.00	92.50	95.00	73.33
User 4	94.00	91.40	93.40	76.67	95.00	80.00
User 5	88.00	96.00	86.60	90.00	80.00	51.43

Table 5. 'Empathy map: thoughts': categories, examples and the assigned scores

Thoughts categories	Designers' empathy map inferences	Users' ratings
The effect of reeds on performance	D1: Choosing the right reed supports the performance.	5
	D2: How could I make reeds that I can be sure will suit my playing needs (e.g., practice, a specific piece at a specific place, etc.)?	5
Environmental effects on reeds and performance	D1: You can keep your best reeds separately, yet their performance depends heavily on circumstances (temperature, humidity).	5
	D2: How can I better protect my reeds and the oboe from changes in the environment (e.g., humidity, temperature, ...), to make sure a good setup stays good longer?	5
Maintenance of the instrument	D1: Cleaning underneath the buttons, cleaning the cushions, should be done more.	4
	D2: Does cleaning every little nook and cranny really make a difference in how the saxophone plays?	4
Adapting to reed demands	D1: Creating the reeds cannot be done all at once, as the wood needs to adapt, better is to do one task per day.	5
	D2: How could I better test my reeds so that I'll know how they'll react when I play them in a specific space with specific acoustics?	5
Music performance	D1: A physical warming up is necessary before beginning to play: muscles and breathing.	5
	D2: How could I better motivate myself to practice properly for every performance?	5

Note: D1 = Designer 1; D2 = Designer 2.

4.2.6. Examples of 'empathy map: thoughts' outcomes

The designers completed the 'think' quadrant of Both and Baggereor's (no date) modified empathy map. The designers synthesised thoughts from each interview and listed them under the 'think' quadrant. The thoughts gathered from all the interviews were grouped into five categories. We present two examples per category and the score given by a musician for a particular thought is presented in Table 5. The musicians were asked to rate the thoughts in terms of how representative were they of their own experiences as users: 1 = *very far from the user's experiences* and 5 = *very close to the user's experiences*.

4.2.7. Examples of 'empathy map: feelings' outcomes

The designers completed the 'feel' quadrant of Both and Baggereor's (no date) modified empathy map. The designers synthesised feelings from each interview and listed them under the 'feel' quadrant. The feelings gathered from all the interviews were grouped into three categories. We present two examples per category and the score given by a musician for a particular feeling is presented in Table 6. The musicians were asked to rate the feelings in terms of how representative were they of their own experiences as users: 1 = *very far from the user's experiences* and 5 = *very close to the user's experiences*.

Table 6. 'Empathy map: feelings': categories, examples and the assigned scores

Feelings categories	Designers' empathy map inferences	Users' ratings
Tediousness regarding instrument cleaning	D1: After a performance she feels tired about having to clean and store the instrument.	5
	D2: Slightly annoyed and bored by the tediousness of cleaning the saxophone and all its parts.	5
Reactions to unpredictability of reeds	D1: Clarinetists shouldn't complain about reeds because oboists have it much harder.	5
	D2: Baffled when she tested a reed just a moment ago and it still plays badly.	5
Music performance	D1: Playing alone is easier than with others due to the multitude of ideas.	5
	D2: Anxiety about the level of precision needed and the number of people relying on you when playing in an orchestra.	5

Note: D1 = Designer 1; D2 = Designer 2.

4.2.8. An example of ideas for improvements

We present four examples out of the 43 ideas for improvements suggested by the designers. The users were also asked to provide a justification for their score.

4.3. Does the designers' empathic accuracy in regard to the musicians positively correlate with design outcomes?

Spearman's correlation analyses between the designers' empathic accuracy scores and their performance on three design outcomes (i.e., empathy map: thoughts, empathy map: feelings, and ideas for improvement) showed medium to large effect sizes. Additionally, the direction of the correlations was sometimes positive and sometimes negative. However, all the correlations were non-significant.

We also explored whether the designers' valence-recognition accuracy (i.e., how correctly they identified whether the emotional tone of a user's entry was positive, neutral or negative) related to their performance in the design outcomes. Spearman's correlation analyses showed lower effect sizes than the ones displayed in Table 8, with the exception of Designer 1's large correlation between valence recognition and ideas for improvement ($rho = -.80$, $p = .10$). As with the previous analysis, the direction of the correlations was sometimes positive and sometimes negative. Similarly, the correlations were all non-significant.

4.4. Does the similarity of the emotional facial expressions of designers and musicians correlate with the designer's empathic accuracy?

We found relatively few instances of frowning (activation of the *corrugator supercilii* muscle) in the dataset, which probably reflects the predominantly positive-valenced emotions experienced by the participants during the interviews. Therefore, we focused our emotional facial expression analysis on smiling (activity of the *zygomaticus major* muscle). A correlation analysis between Designer 1 and

Table 7. Examples of ideas for improvements

Designers' ideas for improvement	User's rating	User's score justification
D1: A cleaning rag that has in one corner a brush that can go in between the knobs, and in another corner a hard element that can keep the knob open and a thinner part that can make it dry. This way no more paper needs to be wasted and fewer cleaning cloths/supports are needed.	3	I understand the idea and it is good but I don't see how it can work … I would need to see the device itself.
D2: Pre-moistened reeds right out of the container, where the container would have a compartment that moistens the end of the reed.	2	Frankly, it sounds a bit gross. Would the reed stay moistened at all times? That would cause an issue with mildew. Also, a reed container with a built-in humidifier already exists. The point of that container isn't to make the reed moist though, but to keep the air in the container humid enough.
D1: The cleaning could be made easier and faster, however it seems to me that the polishing at the end, even though you are tired, is also a ritual to thank your saxophone, as you love it as an extension of yourself. Instead I thus suggest polishing cloths with 'thank you' embroidered in them to constantly remind you why you enjoy to take care of the saxophone, making the ritual more enjoyable.	5	Totally agree, it could be done easier and faster, but it's true that I take care of my saxophone as it was me. Good suggestion for making the ritual more enjoyable. Maybe it can be annoying at the beginning (when you start cleaning your saxophone as a beginner), but actually once you have the routine, it's also part of yourself and your activity. Making music is enjoyable – not only playing music, but everything it concerns around (warming-up, cleaning etc.).
D2: A service that sells cane with known and precise mechanical properties and is priced according to cane quality. You wouldn't just order cane from a supplier, but select a specific quality of cane.	5	This would be the dream of oboe players. We would need some trustful platform that could let us know the best cane at every moment.

Note: D1 = Designer 1; D2 = Designer 2.

the users' *zygomaticus major* signals (117 events), and the empathic accuracy scores obtained from the user (i.e., the empathic accuracy scores reported for all thoughts and feelings across all interviews) did not reveal any relationship (see Figure 2): $p = .51$, $r = -.06$. Similarly, when repeating the same analysis for Designer 2 and the musicians, no correlation was observed: $p = .76$, $r = -.03$. This indicates that similarity in emotional expression during the event was not necessary for (and did not help in) guessing what the musicians were thinking during the events. We also checked if the overall activation levels of either muscle (not their synchrony), of either the designer or the musician, were associated with the empathic accuracy scores, but all these correlations were close to zero as well.

Table 8. The correlation matrix for empathic accuracy scores and design outcomes

	Empathic accuracy	Empathy map thoughts	Empathy map feelings	Ideas for improvements
Empathic accuracy	–			
Empathy map Thoughts	D1: .50 D2: -.40	–		
Empathy map Feelings	D1: -.30 D2: .67	D1: .30 D2: .10	–	
Ideas for improvements	D1: -.70 D2: .60	D1: .20 D2: -.30	D1: .50 D2: .87	–

Note: D1 = Designer 1; D2 = Designer 2.

Table 9. The correlation matrix for valence-recognition accuracy scores and design outcomes

	Valence recognition	Empathy map thoughts	Empathy map feelings	Ideas for improvements
Valence recognition	–			
Empathy map thoughts	D1: -.30 D2: -.10	–		
Empathy map feelings	D1: -.20 D2: .15	D1: .30 D2: .10	–	
Ideas for improvements	D1: -.80 D2: .10	D1: .20 D2: -.30	D1: .50 D2: .87	–

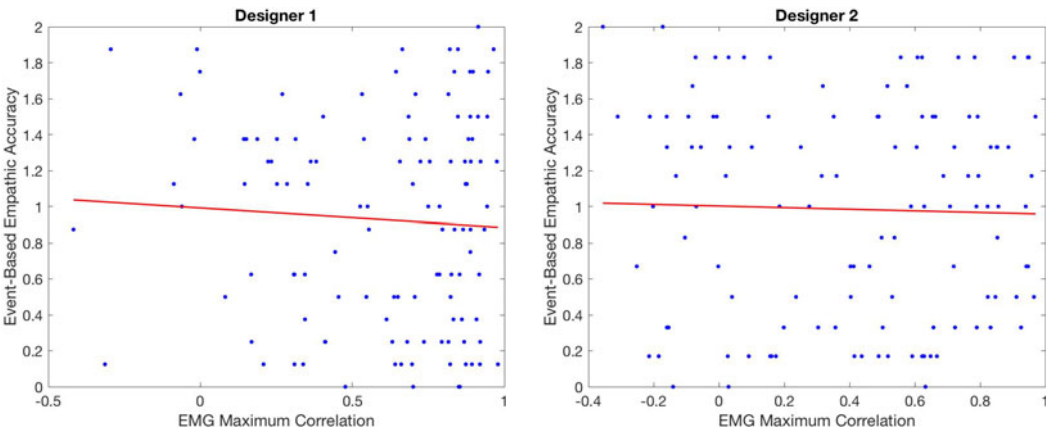*Note:* D1 = Designer 1; D2 = Designer 2.

Figure 2. Scatter plots of the *zygomaticus major* muscle's EMG synchrony and event-based empathic accuracy scores. The blue dots represent the 117 events collected from the five musicians completing the empathic accuracy task. Left: Designer 1; right: Designer 2.

5. Discussion

This study is an initial attempt to rigorously test whether empathy translates into improved design outcomes. We measured empathic accuracy during dyadic interaction in three ways: first, by characterising a designer's empathic accuracy performance; second, by exploring whether the designer's empathic accuracy in regard to the musicians' thoughts and feelings correlates positively with the design outcomes; and thirdly, by exploring whether the similarity of the emotional facial expressions of the designer and users correlated with the designer's empathic accuracy. We found that both designers were capable of correctly identifying about 50% of a user's reported mental content. We obtained small to large correlations between the designers' empathic accuracy and their performance in design outcome tasks, although the contrary direction of the correlations, the lack of statistical significance and the small sample size all limit the interpretation of these results. The analysis of physiological synchrony and empathic accuracy revealed nearly non-existent correlations. Even when based on the performance of just two designers and five users, we collected a considerable amount of data from them; therefore, our results provide important initial information for future research.

5.1. How accurately can the designers understand the group of musicians?

On average, the designers could correctly infer 50% of the mental content reported by five professional musicians. Remarkably, the second designer received similar scores to the interviewer even though he was only exposed to the users through video recordings. The about 50% accuracy obtained by both designers is considerably higher than that found in earlier studies which have reported accuracies from 20% to 30% (Ickes & Hodges 2013; Stueber 2013). For instance, Stinson and Ickes (1992) found that after a casual six-minute interaction, two interacting male strangers had a mean accuracy score of 24% while that of two male friends was 36%. Marangoni *et al.* (1995) found quite similar accuracy scores (23–34%) during psychotherapy sessions. When detecting emotional valence (i.e., whether the inferred thought or feeling had a positive, neutral or negative valence), the designers obtained scores below 50%. Previous studies do not report the participants' correct identification of emotional valence. Thus, it is hard to interpret this result in the light of earlier studies.

The empathic accuracy scores assigned to both designers by two different groups of naive raters were highly reliable and clearly above the minimum standards (.70; Nunnally 1967). Similarly, high reliability values have previously been reported by Ickes (1993) and suggest the suitability of this rating system (Ickes 1993; Ickes 2001) for future studies.

Why then did the designers in our study obtain higher scores than those in previous studies? The musical background of Designer 1 did not seem to give him an advantage over Designer 2. Nor did the progression of interviewing time. The reason for the high-empathic accuracies of both designers may be the semi-structured interview context. The interview had a specific aim of exploring the musicians' experiences with reeds and accessories. Verbal communication was complemented with demonstrations with real objects. In contrast, in unstructured and unexpected conversations (Stinson & Ickes 1992)

and psychotherapy sessions (Marangoni *et al.* 1995), the interviews dealt with more abstract topics and presumably did not include objects that helped one to understand the interviewees' point of view. Thus, in our study the range of the possible mental content of the users was considerably narrower and concrete, making the identification task of the designers easier.

Our results also suggest that these outcomes can be attributed to a concrete design situation in which one is trying to understand a user and not to the trait of empathy in the designers (being more or less empathic). Extensive social psychology research shows that new circumstances (like a designer interviewing a group of musicians about reeds for the first time) have a greater influence on people's behaviour than their trait characteristics (Ross & Nisbett 2011). Similarly, empathic accuracy research suggests that we are faulty judges of our capacity to infer someone else's mental content (Ickes 2003; Stueber 2018), thus it is aligned to previous social psychology research on the influence of specific situations on behaviour. The empathic accuracy method is a performance-based method for measuring the understanding between two or more individuals in a very specific situation. Therefore, it is not a trait measure of a designer.

Designer 1's self-rated empathic accuracy for the dyadic interaction paradigm outcome differed considerably from his actual empathic accuracy. This could be the result of being asked how well he thought he completed the task whereas asking how accurately he inferred each musician's thoughts and feelings would have been more relevant. However, even then a designer would be likely to overestimate her or his actual empathic skills. Previous studies have shown that people have such a tendency (Levenson & Ruef 1992; Ickes & Hodges 2013; Stueber 2013). How would the self-rated empathic accuracy performance differ among professional designers and non-designers? Would the shared educational background of the designers result in higher or lower confidence in their empathic skills when compared to non-designers?

Another relevant finding, although expected, was that the designers' empathic accuracy did not improve over time. Marangoni *et al.* (1995) showed that when respondents to the standard dyadic interaction paradigm were given immediate feedback on the target person's actual thoughts and feelings, there was an increase in empathic accuracy that was not found in a control group that did not get feedback. In the present study, the performance of our designers was not significantly different between the beginning and end of the interview. We wonder whether a designer could increase her or his empathic accuracy towards a user if provided with immediate feedback, thus aiding the designer to understand the context and experience of the user (Kouprie & Sleeswijk Visser 2009; Smeenk, Sturm & Eggen 2017). Future studies could compare the empathic accuracy performance of designers versus non-designers when watching the same contextual interviews and test whether design training translates into differentiated outcomes.

Overall, the dyadic interaction paradigm allows designers to have different insights into users' mental contents. By asking users to report what were they thinking or feeling in great detail and by assigning an emotional valence to this content, designers can have a more precise method with which to trace user experiences. Additionally, the dyadic interaction paradigm allows one to contrast how similar the remembered mental contents of users is to the contents inferred by a designer.

5.2. Does the designers' empathic accuracy in regard to the musicians positively correlate with design outcomes?

It is inconclusive whether the designers' empathic accuracy with regard to the musicians positively correlated with the design outcomes. The designers' empathic accuracy scores, their performance on the empathy map and ideas for improvement tasks showed medium to strong correlations; however, they were completely non-significant. However, different reasons limit their interpretation. The obtained values followed unpredictable directions. Some correlations followed positive trends, as expected, but others had unexpectedly negative correlations. For instance, for Designer 1, the 'think' task of the empathy map had a strong correlation with the empathic accuracy scores and was thus closer to the predicted results. However, this pattern was not found with the 'feel' task of the empathy map. A similar interpretation follows for the outcomes of the ideas for the improvement task. Although the correlation between empathic accuracy and the accurate identification of ideas for improvements was very strong, it was a negative correlation, implying that the higher the empathic accuracy, the lower the accurate identification of ideas for improvements. There are similar difficulties for interpreting the correlations observed in Designer 2 performance. However, with the exception of the 'think' task, his correlations were positive, approaching our prediction. Interestingly, even though his only contact with the users was through videoed interviews, he obtained the same medium to large correlations that Designer 1 did. Perhaps this suggests that a videoed interview can communicate enough information to perform some design tasks.

Another reason that makes it difficult to interpret these effect sizes is possible rating biases and the limitations of the design task surveys responded to by users. Perhaps the musicians were biased when rating the designer whom they most likely knew was the same person who had interviewed them (Dell *et al.* 2012). However, similar high scores were given to Designer 2 (notice the overall high scores obtained by both designers in Table 4) who had no physical contact with the users. Thus, it could simply be that ideas proposed by the experienced designers were genuinely well received by the users. It could also be possible that the design tasks used in this study were problematic. Perhaps choosing only two quadrants from the empathy map deprives it of its full utility. Similarly, the high ratings of the list of ideas for improvement could also be explained by biased users. The discussion remains open regarding how to properly quantify design outcomes. We chose a Likert scale response format to rate design tasks, which is not usually utilised in this way. For example, empathy maps are used as a synthesising and visualisation tool, but it remains unknown to us if users are ever asked to quantify the quality of empathy maps' contents. Although these results suggest the possibility that the dyadic interaction paradigm or the standard stimulus paradigm might not be the best approach to use in order to capture how empathy translates into improved design outcomes, it is too early to draw such a conclusion. Therefore, our assumption that a designer's empathic accuracy performance translates into improved design outcomes must be retested along the lines described in the previous paragraphs.

5.3. Does the similarity of the emotional facial expressions of the designers and musicians correlate with the designers' empathic accuracy?

The similarity of the emotional facial expressions of the designers and users was not at all related to how accurate the designers were in inferring the thoughts or feelings of each of the users in the 117 entries. The negative result could be due to at least two reasons.

The synchronisation of a specific facial muscle (*zygomaticus major*) did not explain empathic accuracy in this study. Our result tentatively suggests that the task of inferring and reporting the thoughts and feelings of others is not helped by prosocial and probably unconscious mirroring of the other's facial expressions. However, this result does not rule out that synchronous facial muscle activity or some other physiological signals could be crucial for empathic accuracy. Previous studies on social interaction indicate that physiology can be used to test synchronisation between individuals and its outcome on different behaviour (Kreibig 2010; Quintana & Heathers 2014; Massaro & Pecchia 2019). For instance, one study concluded that whenever the physiological synchrony (calculated from heart rate and electrodermal activity signals) between subjects was higher, their subjective emotional ratings of a movie they were watching were more similar (Golland *et al.* 2015). Subjects watching the movie were sharing the same space, but did not interact with one another. Therefore, we have to leave open the possibility that synchrony in other physiological signals could reveal an important relevance in relation to understanding others' mind contents (see e.g. Levenson & Gottman 1983; Levenson & Ruef 1992; Zaki *et al.* 2009).

The second reason why facial synchrony did not relate with empathic accuracy scores could be that a strong rapport or the sharing of emotional facial expressions might not be enough to understand the highly specific problems that reed users deal with. Understanding the difficulties related to reeds demands very specialised technical knowledge of acoustics, interpretation, phrasing, reed making, instrument mechanics etc. Perhaps it would be more relevant in the understanding of more emotionally charged topics such as perfectionism or music performance anxiety (Kenny 2011) – topics which can elicit a wider valence and arousal of subjective experiences.

5.4. Limitations and future directions

An evident limitation of the present study is the small number of participants. The low number of musicians interviewed was due to two reasons. First, the measurement session was very long. Every session with a musician took approximately four hours. Despite allowing breaks between sessions, a session was very demanding for the participants. Second, despite efforts to recruit more musicians, only five contacted us. The probable main reason for this is that we aimed to have a very specific group of musicians and thus excluded many that could have been interested in participating. However, the five participants were musicians of very high performing level and thus ideal users for our design problem.

In addition to controlling for the musicians' mood state between the different stages of the study, we should have done the same with the designers. Our main reason for not controlling the designers' mood changes was their comparatively

short participation time (i.e., 1 h 30 min per meeting), distributed across different days, so we reasoned that fatigue would not influence their performance. However, some other factors could have affected their mood and therefore their performance. Therefore, in future studies a more careful control of the mood of all the participants at different stages of the experiment should be done. Other factors which might influence the performance could be, for example, sleeping time, the time of the day and smoking.

We should provide some clarity regarding our implementation of the dyadic interaction paradigm. In a dyadic interaction paradigm, a member of the dyad is asked to infer the mental content of the other member immediately after they have had an interaction. In the present study, we departed from this convention by asking Designer 1 to infer the thoughts and feelings of the users one month after the first interview and three days after the last. We followed this approach because we thought it better to reserve the inference task to the very end. We worried that asking Designer 1 to complete the inference task would have exposed him to a crucial part of the study and prompt him to approach the following interviews differently. It remains open whether Designer 1 would have had higher accuracy scores than Designer 2 if we had closely followed the dyadic interaction paradigm specifications by asking him to infer the users' mental contents right after each interview.

We also believe that communicating our null results is relevant in order to prevent feeding the 'file drawer problem' (Rosenthal 1979) or the higher chance of reporting statistically significant results over null results (Franco, Malhotra & Simonovits 2014). Given the high demands of our method, future studies aiming at adopting it should be informed about its potentialities and limitations.

It is important to discuss some additional lines of future work and other limitations. In this study, we selected an interview as a method of user understanding. However, user understanding is typically created with a wide array of methods – such as multiple interviews, surveys, immersion, iterative prototyping and testing, probes etc. (Sanders & Stappers 2014; Oygür 2018) – instead of only consisting of a one-time interview. As the repeated assumption testing of users has been connected to design success (Häggman, Honda & Yang 2013) and as distinct reactions to user-centred information among designers have been reported (Sugar 2001; Zoltowski et al. 2012), it would be relevant to investigate whether some empathic accuracy paradigm could capture designers' ability or tendency to become more accurate over time. In this study we only tested the empathic accuracy task on the very first interaction between the designers and users. For this particular interaction, and within the imposed limits of our controlled environment, we tried to recreate a real design case by using a contextual interview and capturing the initial design outcomes through two real-world design tools: the empathy map and idea generation. Obviously, the resulting outputs are not a final product or a prototype but rather the first elements for further development. Future work could test how these initial interaction outcomes impact on further design steps or could otherwise look at empathic accuracy over a more comprehensive design process, but this was out of the scope of this study.

Even though we controlled for the English proficiency of the designer and musicians, some of the latter expressed doubts about their language competency. Although all of them were capable of sharing their experiences during the

interview, the language barrier could have hindered the flow of the interview and limited the full expression of the users' experiences and emotions.

6. Conclusion

This study was an initial exploration into quantifying the effect of empathy on design outcomes. The initial results presented here are promising and demonstrate the feasibility of the method. We took two separate approaches to quantifying the designers' understanding of a user. The first one was based on the previous works of Ickes (2001) and Marangoni *et al.* (1995). A relevant finding was that the two designers correctly inferred about half of the five users' stated mental contents. Besides this result, we provided a considerable number of examples in order to illustrate how this method can be used in a design scenario and the type of information that it can provide to researchers. The second approach was based on the work of Levenson and Gottman (1983) and Levenson and Ruef (1992). At the moments that the designers made inferences, their facial muscles were not related to the inference accuracy at the time that the inference was made. But this does not rule out other physiological signals and their potential role as predictors of design outcomes. Given the performance-based nature of the empathic accuracy task, it can be adapted to the very specific circumstances and problematics that designers have to encounter. Therefore, our results encourage future explorations of a method that could expand our understanding of empathy in design based on the measurement of accuracy.

Financial support

Appendix A. Interview model

Preparations

Main theme

Reeds used in woodwind instruments.

Example questions in no particular order

- What has been an excellent reed for you? What made it so good?

- What performances have you done with that reed? How did the reed influence those performances?

- What is a terrible reed you have used? What made it so bad?

- Do you make your own reeds?

- What are important qualities of a reed for you?

- What kind of reeds have you used?

- What was the first reed you ever used?

- Has your preference for certain reed qualities changed over time?

- How long do you use a reed? How long does it typically last? How long does it take to make a reed?

- What was your best performance ever? What reed did you use then?

- When did you start playing and why?

- How did you learn to do your own reeds?

Execution
Introduction (10–15m)
While attaching electrodes

- Welcome

- Introduction interviewer (product designer/researcher)

- Introduction interviewee (which instrument, attributes)

Interview (20–30m)
While electrodes attached and measuring
Subquestions are examples to expand on the stories.

(1) **How long have you been playing [instrument]? Why? 0m**

    (a) What drew you into playing this instrument?

    (b) Have you played any other instruments?

    (c) Were there moments that you played less/more?

(2) **Do you usually play solo or in a group, or several groups? Why? 3m**

    (a) How about when you practice?

    (b) How about when taking lessons?

    (c) How about when performing?

(3) **If you think about preparing to play the [instrument], what do you typically need to do in order to be able to start playing the [instrument]? Why? 10m**

    (a) Do you need to clean, tune, assemble parts?

    (b) What do you need to do after you have finished playing the instrument?

    (c) What is most demanding in relation to being able to play?

(4) **I am actually quite interested in these reeds. If you think about the reeds you use, what makes one stand out for you, what makes it good? Why? 15m**

    (a) Do you make your own, or have a special supplier?

    (b) What other kinds of reeds have you used?

    (c) Has your preference of reeds changed over time?

    (d) What have been some good and bad experiences for you with reeds?

(5) **If you think about your performances, what are your more memorable performances? 21m**

(a) What has been an enjoyable performance for you?

(b) What made that an enjoyable performance?

(c) What has been a less enjoyable performance for you?

(d) What made that performances less enjoyable?

(e) If you think about these performances, did the reed influence the enjoyability of that performance?

(6) **I know this has been short, but we're almost running out of time, so let's go to the last question. Do you have any other experiences with your instrument that you would like to share? 28m**

Closing (10–15m)
While detaching electrodes

• Thank you. Do you have any questions or comments for me?

Appendix B. Filling in Thoughts or Feelings you Remembered

You will now rewatch the interview. Please, stop the recording at those points where you remembered having had a specific thought or feeling. Remember, you are asked to write down thoughts and feelings you remembered instead of new thoughts or feelings that you might have while rewatching the interview.

Under the column 'time' indicate the specific time on the recording where you remembered those thoughts or feelings. Report all of the thoughts and feelings you remember having as accurately, honestly and completely as possible under the 'thought or feeling' column. Please, use a different box for each thought or feeling you report. Finally, choose the tone of the emotion you experienced when remembering a specific thought or feeling:

• Positive: $+$

• Neutral: 0

• Negative: $-$

Example of how to record your answers

TIME	THOUGHT OR FEELING	$+, 0, -$
5.36	I was: thinking about my instrument	$\oplus$ 0 $-$
7.12	I was: feeling uncomfortable in this chair	$+$ 0 $\ominus$
7.57	I was: thinking of that music lesson	$+$ $\textcircled{0}$ $-$

After you have completed the task or at any point thereafter you will be allowed to delete any thought or feeling entry and any portion of the video recording that you would prefer remain private.

Thank you very much!

Appendix C. Filling in your Inferred Thoughts or Feelings

You will be showed the interview between the musician and you. Please, read through the following instructions. The video will be automatically paused by the researcher. Every time the video is paused, you are to write down what you think the musician was thinking or feeling at that moment by filling in one of these slots. Please, use a different box for each thought or feeling you inferred. Remember, your task is to make a straightforward inference about what the musician was actually thinking or feeling at each of the stop points on the video. Once you have written your answer, press the space bar to continue and repeat the process every time the video pauses. Finally, choose the tone of the emotion you think she experienced when having a specific thought or feeling:

- Positive: $+$
- Neutral: 0
- Negative: $-$

Example of how to record your inferences

TIME	THOUGHT OR FEELING	+, 0, −
5.36	She was: thinking about her instrument	⊕ 0 −
7.12	She was: feeling uncomfortable in her chair	+ 0 ⊖
7.57	She was: thinking about that music lesson	+ ⓪ −

Thank you very much!

Appendix D. Example of User Response Sheet

TIME	THOUGHT OR FEELING	+, 0, −
	I was:	+ 0 −
	I was:	+ 0 −
	I was:	+ 0 −
	I was:	+ 0 −
	I was:	+ 0 −

Appendix E. Example of Designer Response Sheet

TIME	THOUGHT OR FEELING	+, 0, −
	He/she was:	+ 0 −
	He/she was:	+ 0 −
	He/she was:	+ 0 −
	He/she was:	+ 0 −
	He/she was:	+ 0 −

Appendix F. Instructions to Rate Similarity Between Thoughts and Feelings

Your task is to compare the written content of the 'actual thoughts or feelings' column with those of the 'inferred thoughts or feelings' one. Please, rate how similar do you think they are in terms of content by using the following scale:

2 = essentially the same content.

1 = somehow similar, but not the same content.

0 = essentially different content.

Next you will see 6 examples, two per each scoring point. Should you have any questions, please contact the researcher. Thank you!

Actual thoughts or feelings	Inferred thoughts or feelings	How similar are they? (Max = 2, Min = 0)
I was feeling silly because I couldn't remember my teacher's name.*	Feeling sorta odd for not remembering her teacher's name.*	2
Feeling very hungry.	She was feeling hungry.	2
I was thinking that I was not missing anything I didn't want to miss. I was thinking that I came to school to learn, not to join organizations.*	She was thinking about what he was missing in school.*	1
That song I need to practice.	She was thinking about music.	1
I was thinking about a previous production of the play in another city that a local radio personality was in.*	She was thinking if I would ask her out.*	0
I was thinking about my daughter's upcoming birthday.	About the new theater play.	0

*Examples original from Ickes, W. (2001) Measuring empathic accuracy in Hall, J. A. and Bernieri F. J. (eds.) *Interpersonal Sensitivity: Theory and Measurement.* Erlbaum, Mahwah, N. J.: Erlbaum, pp. 219–241.

Example of an Actual Rating Case

Actual thoughts or feelings	Inferred thoughts or feelings	How similar are they? (Max = 2, Min = 0)
I was: curious about what the interview was going to be about.	He/she was: feeling slightly nervous about the interview.	2
I was: amused by the joke.	He/she was: feeling entertained about the interviewer's comment of not being able to design an instrument.	2
I was: excited to begin since I like answering questions.	He/she was: thinking back of when she first started playing.	0
I was: slightly bored, I've told this story a million times.	He/she was: feeling slightly uncomfortable about sharing a personal story from her past.	1
I was: amused by the fact that I really had no idea what the clarinet was at the time.	He/she was: feeling funny about not knowing the clarinet when beginning to play.	2
I was: thinking that I've mentioned my mom so many times it might sound like she made all my decisions.	He/she was: thinking back of how her mother influenced the decision of choosing the clarinet.	2

Appendix G. Empathy Map and Ideas for Improvements Tasks

Create an empathy map to summarise and synthesise the key insights you came up with after watching the interview. This empathy map should have two columns called think and feel. Imagine you are completing this task as part of a professional task.

Feel free to work with the materials provided for you, but please write down your answers on this computer after completing the task.

Additionally, write down which are the most important features you came up with for the instrument after having watched the interview.

Think *What might your user be thinking? What does this tell you about her beliefs?*	**Feel** *What emotions might your subject be feeling?*

References

Almeida, A., **George, D.**, **Smith, J.** & **Wolfe, J.** 2013 The clarinet: How blowing pressure, lip force, lip position and reed 'hardness' affect pitch, sound level, and spectrum. *J. Acoust. Soc. Am.,* **134** (3).

Baron-Cohen, S. & **Wheelwright, S.** 2004 The Empathy Quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders* **34**, 163–175.

Bird, G. & **Viding, E.** 2014 The self to other model of empathy: providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience and Biobehavioral Reviews* **47**, 520–532.

Battarbee, K., **Baerten, N.**, **Hinfelaar, M.**, **Irvine, P.**, **Loeber, S.**, **Munro, A.** & **Pederson, T.** 2002 Pools and satellites: Intimacy in the city. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, pp. 237–245. ACM Press.

Both, T. & **Baggereor, D.** *Bootcamp Bootleg* [Online]. Available at: https://dschool.stanford.edu/resources/the-bootcamp-bootleg (Accessed : 15 July 2019). d. School Hasso Plattner Institute of Design at Stanford.

Brown, T. 2009 *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation*. HarperBusiness.

Carr, L., **Iacoboni, M.**, **Dubeau, M. C.**, **Mazziotta, J. C.** & **Lenzi, G. L.** 2003 Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences* **100** (9), 5497–5502.

Cuff, B. M. P., **Brown, S. J.**, **Taylor, L.** & **Gowat, D. J.** 2016 Empathy: a review of the concept. *Emotion Review* **8** (2), 144–153.

Davis, M. H. 1980 A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* **10**, 85; Available at: http://www.ucp.pt/site/resources/documents/ICS/GNC/ArtigosGNC/AlexandreCastroCaldas/24_Da80.pdf.

Dell, N., **Vaidyanathan, V.**, **Medhi, I.**, **Cutrell, E.** & **Thies, W.** 2012 Yours is better!: Participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1321–1330. ACM Press.

Ellis, P. D. 2010 *The essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press.

Ellis, J., **Schroder, H.**, **Patrick, C.** & **Moser, J.** 2017 Emotional reactivity and regulation in individuals with psychopathic traits: evidence for a disconnect between neurophysiology and self-report. *Psychophysiology* **54** (10), 1–12.

Franco, A., **Malhorta, N.** & **Simonovits, G.** 2014 Publication bias in the social sciences: unlocking the file drawer. *Science* **345** (6203), 1502–1505.

FSenSync (Version 101) [Computer software]. Finland: Förger Analytics.

Gates, K. M., **Gatzke-Kopp, L. M.**, **Sandsten, M.** & **Blandon, A. Y.** 2015 Estimating time-varying RSA to examine psycho-physiological linkage of marital dyads. *Psychophysiology* **52**, 1059–1065.

Ghosh, D., **Olewnik, A.**, **Lewis, K.**, **Kim, J.** & **Lakshmanan, A.** 2017 Cyber-empathic design: A data-driven framework for product design. *J. Mech. Des.* **139** (9), 1–12.

Golland, Y., **Arzouan, Y.** & **Levit-Binnun, N.** 2015 The mere co-presence: synchronization of autonomic signals and emotional responses across co-present individuals not engaged in direct interaction. *PLoS One* **10** (5).

Golland, Y., **Hakim, A.**, **Aloni, T.**, **Schaefer, S.** & **Levit-Binnun, N.** 2018 Affect dynamics of facial EMG during continuous emotional experiences. *Biological psychology* **139**, 47–58.

Hess, J. L. & **Fila, N. D.** 2016 The manifestation of empathy within design: findings from a service-learning course. *CoDesign* **1–2**, 93–111.

Hess, J. L., **Strobel, J.** & **Pan, R.** 2016 'Voices from the workplace: practitioners' perspectives on the role of empathy and care within engineering. *Eng. Stud.* **8** (3), 212–242.

Hess, J. L., **Strobel, J.**, **Pan, R.** & **Wachter Morris, C. A.** 2017 Insights from industry: a quantitative analysis of engineers perceptions of empathy and care within their practice. *Eur. J. Eng. Educ.* **42** (2), 1128–1153.

Häggman, A., **Honda, T.** & **Yang, M. C.** 2013. The influence of timing in exploratory prototyping and other activities in design projects, *Proceedings of the ASME 2013 IDETC/CIE, Charlotte, North Carolina*. Available at: doi:10.1115/DETC2016-59417.

Ickes, W. 1993 Empathic accuracy. *Journal of Personality* **61** (4), 587–610.

Ickes, W. 2001 Measuring empathic accuracy. In *Interpersonal Sensitivity: Theory and Measurement* **(ed. J. A. Hall & F. J. Bernieri)**, pp. 219–241. Erlbaum, Erlbaum, Mahwah, N. J.

Ickes, W. 2003. *Everyday Mind Reading: Understanding What Other People Think and Feel*. Prometheus Books.

Ickes, W., **Bissonette, V.**, **Garcia, S.** & **Stinson, L.** 1990 Implementing and using the dyadic interaction paradigm. In *Review of Personality and Social Psychology: Volume 11*, Research Methods in Personality and Social Psychology, pp. 16–44. Sage.

Ickes, W. & **Hodges, S. D.** 2013 Empathic accuracy in close relationships. In *Oxford Library of Psychology. The Oxford Handbook of Close Relationships* **(ed. J. A. Simpson & L. Campbell)**, pp. 348–373. Oxford University Press.

Ickes, W., **Stinson, L.**, **Bissonnette, V.** & **Garcia, S.** 1990 Naturalistic social cognition: empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology* **59** (4), 730–742.

IDEO.org 2015. *The Field Guide to Human-Centered Design 2015* [Online]. Available at: http://www.designkit.org/resources/1 (Accessed: 23 May 2018). IDEO.org.

Jackson, P. L., **Meltzoff, A. N.** & **Decety, J.** 2005 How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage* **24**, 771–779.

Kankainen, A., **Vaajakallio, K.**, **Kantola, V.** & **Mattelmäki, T.** 2012 Storytelling Group – a co-design method for service design. *Behaviour & Information Technology* **31** (3), 221–230.

Kenny, D. T. 2011 *The Psychology of Music Performance Anxiety*. Oxford University Press.

Kleinbub, J. R. 2017 State of the art of interpersonal physiology in psychotherapy: a systematic review. *Frontiers in Psychology* **8**.

Köppen, E. & **Meinel, C.** 2015 Empathy via design thinking: creation of sense and knowledge. In *Design Thinking Research* **(ed. H. Plattner, C. Meinel & L. Leifer)**, pp. 15–28. Springer International Publishing.

Koskinen, I. & **Battarbee, K.** 2003 Introduction to user experience and empathic design. In *Empathic Design, User Experience in Product Design* **(ed. I. Koskinen, K. Battarbee & T. Mattelmäki)**, pp. 37–50. IT Press.

Kouprie, M. & **Sleeswijk Visser, F.** 2009 A framework for empathy in design: Stepping into and out of the user's life. *Journal of Engineering Design* **20** (5), 437–448.

Kramer, J., **Agogino, A. M.** & **Roschuni, C.** 2016. Characterizing competencies for human-centered design, *Proceedings of the ASME 2016 IDETC/CIE, Charlotte, North Carolina*. Available at: doi:10.1115/DETC2016-60085.

Kreibig, S. D. 2010 Autonomic nervous system activity in emotion: A review. *Biological Psychology* **84**, 394–421.

Lapate, R. C., **Van Reekum, C. M.**, **Schaefer, S. M.**, **Greischar, L. L.**, **Norris, C. J.**, **Bachhuber, D. R.** & **Davidson, R. J.** 2014 Prolonged marital stress is associated with short-lived responses to positive stimuli. *Psychophysiology* **51** (6), 499–509.

Ledet, D. A. 1981 *Oboe Reed Styles, Theory and Practice*. Indiana University Press.

Levenson, R. W. & **Gottman, J. M.** 1983 Marital interaction: physiological linkage and affective exchange. *Journal of Personality and Social Psychology* **45** (3), 587–597.

Levenson, R. W. & **Ruef, A. M.** 1992 Empathy: a physiological substrate. *Journal of Personality and Social Psychology* **63** (2), 234–246.

Light, S. N., **Moran, Z. D.**, **Swander, L.**, **Le, V.**, **Cage, B.**, **Burghy, C.**, **Westbrook, C.**, **Greishar, L.** & **Davidson, R. J.** 2015 Electromyographically assessed empathic concern and empathic happiness predict increased prosocial behavior in adults. *Biol Psychol.* **104**, 116–129.

Lin, J. & **Seepersad, C. C.** 2007. Empathic lead users: the effects of extraordinary user experiences on customer needs analysis and product redesign, *Proceedings of the ASME 2007 IDETC/CIE, Las Vegas, Nevada*. Available at: doi:10.1115/DETC2007-35302.

Lunkenheimer, E., **Tiberio, S. S.**, **Buss, K. A.**, **Lucas-Thompson, R. G.**, **Boker, S. M.** & **Timpe, Z. C.** 2015 Coregulation of respiratory sinus arrhythmia between parents and pre- schoolers: differences by children's externalizing problems. *Developmental Psychobiology* **57**, 994–1003.

Marangoni, C., **Garcia, S.**, **Ickes, W.** & **Teng, G.** 1995 Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology* **68** (5), 854–869.

Marci, C. D., **Ham, J.**, **Moran, E.** & **Orr, S. P.** 2007 Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *Journal of Nervous and Mental Disease* **195**, 103–111.

Massaro, S. & **Pecchia, L.** 2019 Heart rate variability (HRV) Analysis: a methodology for organizational neuroscience. *Organizational Research Methods* **22** (1), 354–393.

Moreira, D., **Azeredo, A.** & **Barbosa, F.** 2019 Neurobiological findings of the psychopathic personality in adults: one century of history. *Aggression and Violent Behavior* **47**, 137–159.

Nagel, J. 2010 Treatment of music performance anxiety via psychological approaches: a review of selected CBT and psychodynamic literature. *Medical Problems of Performing Artists* **25**, 141–148.

Nunnally, J. C. 1967 *Psychometric Theory*. McGraw Hill.

Oygür, I. 2018 The machineries of user knowledge production. *Design Studies* **54**, 23–49.

Palumbro, R. V., **Marraccini, M. E.**, **Weyandt, L. L.**, **Wilder-Smith, O.**, **McGee, H. A.**, **Liu, S.** & **M. S., Goodwin** 2017 Interpersonal autonomic physiology: a systematic review of the literature. *Personality and Social Psychology Review* **21** (2), 99–141.

Pang, M. A. & **Seepersad, C. C.** 2016. Crowdsourcing the evaluation of design concepts with empathic priming, *Proceedings of the ASME 2016 IDETC/CIE, Charlotte, North Carolina*. Available at: doi:10.1115/DETC2016-59417.

Preston, S. D. & **de Waal, F. B. M.** 2002 Empathy: its ultimate and proximate bases. *Behavioral and Brain Sciences* **25**, 1–72.

Postma, C. E., **Zwartkruis-Pelgrim, E.**, **Daemen, E.** & **Du, J.** 2012 Challenges of doing empathic design: experiences from industry. *International Journal of Design* **6** (1), 59–70.

Quintana, D. S. & **Heathers, J. A.** 2014 Considerations in the assessment of heart rate variability in biobehavioral research. *Frontiers in Psychology* **5**, 1–10.

Rasoal, C., **Danielsson, H.** & **Jungert, T.** 2012 Empathy among students in engineering programmes. *Eur. J. Eng. Educ.* **37** (5), 427–435.

Raviselvam, S., **Anderson, D.**, **Hölttä-Otto, K.** & **Wood, K. L.** 2018. Systematic framework to apply extraordinary user perspective to capture latent needs among ordinary users, *Proceedings of the ASME 2018 IDETC/CIE, Quebec City, Quebec, Canada*. Available at: doi:10.1115/DETC2018-86263.

Raviselvam, S., **Sanaei, R.**, **Blessing, L.**, **Hölttä-Otto, K.** & **Wood, K. L.** 2017. Demographic factors and their influence on designer creativity and empathy evoked through user extreme conditions, *Proceedings of the ASME 2017 IDETC/CIE, Cleveland, Ohio*. Available at: doi:10.1115/DETC2017-68380.

Rosenthal, R. 1979 The 'file drawer problem' and tolerance to null results. *Psychological Bulletin* **86** (3), 368–641.

Ross, L. & **Nisbett, R. E.** 2011 *The Person and the Situation: Perspectives of Social Psychology*. McGraw-Hill.

Sanders, E. B. N. & **Stappers, P. J.** 2014 Probes, toolkits and prototypes: three approaches to making in codesigning. *CoDesign* **10** (1), 5–14.

Singer, T., **Seymour, B.**, **O'doherty, J.**, **Kaube, H.**, **Dolan, R. J.** & **Frith, C. D.** 2004 Empathy for pain involves the affective but not sensory components of pain. *Science* **303** (5661), 1157–1162.

Soto, J. A. & **Levenson, R. W.** 2009 Emotion Recognition across Cultures: the Influence of Ethnicity on Empathic Accuracy and Physiological Linkage. *Emotion* **9** (6), 874–884.

Shamay-Tsoory, S. G. 2011 The neural bases for empathy. *The Neuroscientist* **17** (1), 18–24.

Smeenk, W., **Sturm, J.** & **Eggen, B.** 2017. Empathic handover: How would you feel? Handing over dementia experiences and feelings in empathic co-design, *CoDesign*. Available at: doi:10.1080/15710882.2017.1301960.

Smeenk, W., **Tomico, O.** & **Van Turnhout, K.** 2016 A systematic analysis of mixed perspectives in empathic design: not one perspective encompasses all. *Int. J. Des.* **10** (2), 31–48.

Strobel, J., **Hess, J.**, **Pan, R.** & **Wachter Morris, C. A.** 2013 Empathy and care within engineering: qualitative perspectives from engineering faculty and practicing engineers. *Eng. Stud.* **5** (2), 137–159.

Stueber, K. 2018. Empathy in E. N. Zalta (ed.) *The Stanford encyclopedia of philosophy* [Online] Available at: https://plato.stanford.edu/archives/spr2018/entries/empathy/.

Sugar, W. A. 2001 What is so good about user-centered design? Documenting the effect of usability sessions on novice software designers. *Journal of Research on Computing in Education* **33** (3), 235–250.

Surma-aho, A., **Björklund, T.** & **Hölttä-Otto, K.** 2018. Assessing the development of empathy and innovation attitudes in a project-based design thinking course, *Research presented at the ASEE Annual Conference and Exposition, Salt Lake City, Utah*. Available at: https://www.asee.org/public/conferences/106/papers/21511/view.

Thompson, S. C. 1979 The effect of the reed resonance on woodwind tone production. *The Journal of the Acoustical Society of America* **66**, 1299–1307.

van Reekum, C. M., **Schaefer, S. M.**, **Lapate, R. C.**, **Norris, C. J.**, **Greischar, L. L.** & **Davidson, R. J.** 2010 Aging is associated with positive responding to neutral

information but reduced recovery from negative information. *Social Cognitive and Affective Neuroscience* **6** (2), 177–185.

Vaughan, M. R., **Seepersad, C. C.** & **Crawford, R. H.** 2014. Creation of empathic lead users from non-users via simulated lead user experiences, *Proceedings of the ASME 2014 IDETC/CIE, Buffalo, New York*. Available at: doi:10.1115/DETC2014-35052.

Walther, J., **Miller, S. E.** & **Sochacka, N. W.** 2017 A model of empathy in engineering as a core skill, practice orientation, and professional way of being. *J. Eng. Educ.* **106** (5), 123–148.

Watson, D., **Clark, L. A.** & **Tellegen, A.** 1988 Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology* **54** (6), 1063–1070.

Wong, K., **Norris, R. L.**, **Siddique, Z.**, **Altan, C.**, **Baldwin, J.** & **Merchan-Merchan, W.** 2016. Cognitive empathy in design course for a more inclusive mechanical engineering, *Proceedings of the ASME 2016 IDETC/CIE, Charlotte, North Carolina*. Available at: doi:10.1115/DETC2016-60382.

Zaki, J. & **Ochsner, K. N.** 2012 The neuroscience of empathy: progress, pitfalls and promise. *Nature Neuroscience* **15** (5), 675–680.

Zaki, J., **Weber, J.**, **Bolger, N.** & **Ochsner, K. N.** 2009 The neural basis of empathic accuracy. *Proceedings of the National Academy of Sciences of the United States of America* **106** (27), 11382–11387.

Zoltowski, C. B., **Oakes, W. C.** & **Cardella, M. E.** 2012 Students' ways of experiencing human-centered design. *Journal of Engineering Education* **101** (1), 28–59.