This manuscript, first authored by Xiao Ge, is under finalization for submission in December 2020 to a design research journal: *Design Studies*

Situated Emotion: How to Measure The "Oh no", "Oh phew" and "Oh yay" of Design Thinking in the Wild

Author(s) information

Abstract: Design is filled with rich emotional experiences from resolving team conflict to breaking impasse to hitting the big eureka. How to capture momentary emotional responses in the wild and across time? We present a multimodal study where designers' emotion is identified by triangulating bodily signals — vocal pitch and electrodermal activity — with multi-angle video analysis and retrospective self-report analysis. Through this mixed-methods study of experienced designers in naturalistic settings, we develop a novel methodological approach to measure and analyze situated emotion. Specifically, we promote the notion of conditional concordance to deal with discordance between different emotion measures. We demonstrate its methodological value in identifying, characterizing and understanding a designer's pronounced emotional experiences, as well as assessing intraindividual change, and interpersonal and intergroup differences. Together, the research highlights the importance to study designer emotion which has been overshadowed by the focus on design cognition and behavior in design research.

Keywords: mixed methods; designer emotion; time-series data analysis; physiological measure; team-based engineering and design

Highlights:

- Novel, rigorous mixed-methods study of experienced designers' emotion in situ
- Identification and analysis of involuntarily leaked surprise, confusion and interest
- Emotional fluctuation covary with change of design phase and social situation
- Interindividual and intergroup comparison of emotion dynamics

Design thinking, a phrase that is widely believed to carry the core of design ability, is often criticized for its lack of doing (Micheli, et al, 2019). But there is another missing block, which is design feeling (Coyne, 2015). “We are not thinking machines. We are feeling machines that think”, as Antonio Damasio asserts in *Descartes' error* (2006). A key to understanding designers’ behavior and ability is emotion (Balters & Steinert, 2017). Most studies of designers, however, have only focused on their behavioral and cognitive characteristics (Casakin & Goldschmidt, 1999; Chiu, 2003; Ahmed & Wallace, 2004; Cross, 2004; Yilmaz & Seifert, 2011; Vallet, et al., 2013; Chai, et al., 2015). Far less work has empirically examined the role of affect in design (Sas & Zhang, 2010; Gerber & Carroll, 2012). The traditional view of technical rationality (Schön, 1983; Harris, 1983) still dominates some worlds of design, such as in engineering. Emotion is a neglected dimension of engineering professional identity. Understanding the role of emotion is arguably even more important as what face designers are under-defined problems, unstructured objectives, uncertain processes and unfamiliar stakeholders, which are now the new norms of design in the changing world (Dorst, 2015).

Despite the lack of empirical studies, design scholars have long argued that emotion plays a key role in design. Many products come into being out of designers’ frustration and dissatisfaction with the status quo. This is exemplified by Sam Farber’s OXO Good Grips, which was inspired by Farber’s wife who had trouble holding a peeler due to arthritis (Bennett, et al., 2019). John Arnold argued that it takes a fervent emotion, a “daring spirit”, to fight for what the designer thinks is right (Arnold, in Clancey, 2016). While emotion can guide creative efforts, “emotional blocks” need to be overcome to avoid fixation and gain creative insights (Maslow, Arnold, in Clancey, 2016; Crilly, 2019). In *The Reflective Practitioner*, Donald Schön (1983) asserts that moments of surprise and confusion can function as springboards for reframing perspectives. The recognition of a creative insight is regarded as a “highly emotional step” by Kees Dorst and Nigel Cross (2001). During the design process, a big source of energy and stress in design is working in teams (Kilker, 1999; Dym & Little, 1999; Jung, 2016; Paletz, et al., 2011). And once the design solution is finalized, designers often need to have enough emotional energy to carry the solution through bureaucracy, skepticism and resistance to become a real-life product (Christensen, 2013).

The current research aims to uncover the internal experiences underlying such a design journey that inevitably encompasses emotional ups and downs. We focus on addressing the methodological dilemma posed by the elusive nature of internal experiences. In technology-equipped lab studies, designers typically work in social isolation from users and other key stakeholders, whereas field studies are not technology-friendly, making continuous monitoring of emotion a methodological challenge. Even with the ubiquitous computing (e.g., with smartphones and wearables), how to capture the latent variables of emotion continues to be a challenge (Mauss & Robinson, 2009; Ram, et al., 2017).

To overcome the challenges, we have created a multimodal dataset where dyads of experienced designers worked on a design challenge, each for half a day in a natural environment. Given the non-negotiable priority of ecological validity for this research, we experimented with several unobtrusive objective measures, including speech acoustics and electrodermal activity, and compared them with multi-angle video-based observational data as well as designers' retrospective self-report. The purpose of such a multimodal dataset is to equip the investigation with multiple perspectives, and also to examine content validity through triangulation, as is the golden standard in both affective science through correlation analysis (Mauss, et al, 2004) and mixed-methods study that involve qualitative results (Jick, 1979; Valentine, et al., 2015).

We will show that the study result demonstrates the capability of this mixed-methods approach to capture the emotion-in-action of designers situated in natural design contexts. Together, our research signifies the importance of emotion research in understanding designer ability and strengthens the methodological foundations of design research of emotion. We call for more research of emotion to augment our understanding of design expertise and ability.

1. How is emotion measured in affective science?

Mood, affect and emotion research of professionals and their group work casts a wider net outside design research, such as in psychology (Csikszentmihalyi, 2013; Davis, 2009; Gino & Ariely, 2012) and management and sociological research (Barsade, 2002; Amabile, et al., 2005; Kunda, 2009). Emotion research is also well established in developmental psychology (Ram, et al., 2011; Lougheed, et al., 2020) and learning sciences (Op't Eynde & Turner, 2006; Pekrun, et al., 2014). Despite the long history of affective science, how to measure emotion is one of the most vexing problems scholars face (Mauss & Robinson, 2009; Barrett, 2017). The elusive nature of emotion is reflected by the lack of common grounded theories and conceptualizations – divergence in what to measure (e.g., discrete states or emotional dimensions) and how to measure (e.g., autonomic, behavioral, or brain states).

There are two mainstream conceptions of emotion. The first one - emotional specificity perspective suggests that emotional states, such as anger, sadness, contempt, are each experientially, physiologically and behaviorally distinctive. Measures that rely on this perspective distinguish emotional states based on distinctive behavioral patterns, such as facial expression analysis. However, it would erroneously link a laughing facial expression to joy when laughter can be driven by anger, ecstasy, or nervousness, depending on the context. This perspective, though aligns with layperson beliefs in some cultures, has reported more inconsistent results, faced more challenges against scientific evidences (Barrett, 2017), and failed more often when characterizing people from non-Western cultures (Kitayama & Markus, 2000; Mesquita & Kawasaka 2002).

The dimensional perspective, in contrast, better accounts for the complexity of emotional behaviors, across demographic dimensions and social situations. Here, emotion is diagrammed by several primitive and universal dimensions, such as valence, arousal, dominance, and situational content (Russell, 1980; Barrett, et al., 2007). Core emotion is mapped by a two-by-two framework with one axis being pleasure-displeasure (valence) and the other — activation-deactivation (arousal). The dimensional perspective is perceived as the more scientific approach to emotion (Barrett, 2017).

Emotional arousal has been made easier to measure objectively by technological advances in accessing physiological responding, such as blood pressure, cardiac output, heart rate, pupil size (Bradley, 2008), electrodermal activity (or skin conductance response, Boucsein, 2012), as well as speech acoustics (Voigt, et al., 2014). Electrodermal activity is the most widely applied psychophysiological response system (Dawson, 2017), and can be made nonobtrusive using wireless wearables such as Empatica E4. Vocal measures are also nonobtrusive. Amongst voice characteristics, voice pitch is the most reliable measure of arousal (Mauss & Robinson, 2009).

In comparison to arousal, objective measure of valence has more limited choices. Startle response magnitude measured by eye blink is found to be a robust measure of valence in early emotion research, but it only operates in well-defined experimental conditions. Another reliable indicator of valence is facial behavior (Russel, 1994), which can be reliably computed via algorithms (Martinez, et al., 2017). It is worth noting that facial behavior analysis is more often utilized for identifying emotional states, the other perspective that receives much controversy. Emotion detection through written text (e.g., twitter, email) is also a widely adopted approach (Kahn, et al., 2007; Calvo & D'Mello, 2010; Vosoughi, et al., 2018). However, evidence is scarce regarding the accuracy of oral content-based emotion analysis. A more comprehensive review of methods to measure valence and arousal can be found elsewhere (Mauss & Robinson, 2009; Calvo & D'Mello, 2010).

Subjective measures of emotion include observer's report (Bartel & Saavendra, 2000; Barsade, 2002; Tsai, et al., 2006) and self-report. Self-report of emotion is typically used in interviews (Sas & Zhang, 2010), reflective journal (Hariharan, 2011), experience sampling (Pychyl, et al., 2000; Csikszentmihalyi & Larson, 2014), and surveys where participants are asked to rate their emotions based on a set of scales retrospectively (Barsade, et al., 2002; Todorova, et al., 2014) or momentarily (Barrett, 1997). Problems such as social desirability bias, accessibility and consistency compromise the validity and reliability of subjective measures (Ram, et al., 2017).

Apparently, emotion cannot be measured by any single method considered alone (Lang, 1988). Any one measure of emotion has its own limitations and biases and is likely associated with variance unique to it (Mauss & Robinson, 2009). In previous work, different physiological, cognitive, expressive and behavioral measures of emotion are found to be poorly correlated or uncorrelated at all, showing little

support to the emotional response concordance view (Cacioppo et al., 2000; Lang, 1988; Mauss et al., 2004). On the other hand, it was shown that concordance is contingent on emotion intensity and how pronounced the cognitive element of the emotion is (Mauss et al., 2004; Reisenzein, 2000). Integrative analysis that combines different emotional measures has proven to be more effective in some studies (Kessous, et al., 2010; Poria, et al., 2017; D' Mello & Kory, 2015).

2. How is emotion researched in design?

Within the limited research work of designers' own emotions, most applied self-report measures, while a handful others took advantage of spoken words and bodily signals. But even with self-report measures, design research of emotion only embarked after the 2000s. In an interview study with expert designers, Sas & Zhang (2010) identified emotions during the process of creative problem solving (e.g., incubation stage is characterized by impasse) and how emotions are regulated for optimal performance and used as an intuitive guide for decision-making. On the same research topic, Hutchinson (2018) explored how graphic designers describe and conceptualize their emotional experiences during ideation by applying phenomenography on a set of interview data. In Gerber & Carroll (2012), an ethnographic study was conducted to inductively explore designers' feelings when their work is facilitated by low-fidelity prototyping. Using a similar approach, Ge & Leifer (2020) explored the emotional journey of how novice designers get stuck and get unstuck through perplexing experiences within design. Last but not least, Hu, et al. (2010) used survey, a mainstream approach outside the design research community, to test how engineers' affective states were correlated to task difficulty.

To increase the self-report assessment granularity, Safin, et al. (2016) asked subjects to self-rate their emotional states using a self-rating tool while self-observing through a video recording. The method proposed in Safin et al. (2016) shows potential of mapping designers' emotional pattern during ideation. Similarly, Jung (2016) asked subjects to self-rate their emotional valence through a scoring knob as they watched themselves in a video recording right after completing the team-based design activity. With that, Jung found that a balance of positive and negative affect as well as hostile affect, as measured by subjects' self-rating, predicts design teams' future performance. This method achieves higher granularity than interview study, but its validity could suffer from the same limitations of interview studies. In addition, it would not be practical to have subjects manually assess experiences of a longer, more realistic time span.

Studies of emotion that capitalize on body language are scarce but emerging in design research. Behoora & Tucker (2015) took this very different methodological stance and assessed designers' emotional states through body language and machine learning. Another study conducted by Jung (2016) analyzed emotional responses in time series data, using Specific Affects Coding System (SPAFF), or

more specifically, a combination of facial muscle movement, speech prosody, verbal content and body posture. Zhou, et al. (2019) applied facial behavior-based emotion analysis to explore how individual design work and group design work differ in terms of emotion.

Similar to body language, transcript-based proxy for emotional assessment is also advantageous for its objectiveness. By applying linguistic analysis on design meeting transcripts, Dong, et al. (2009) found the positive/negative appraisals are associated with different types of knowledge generation and integration. Also using semantic analysis, Ewald, et al (2019) studied how group emotional valence differs across design thinking stages. One methodological challenge it faces is that designers may censor true feelings from verbal communication in social settings. Besides, the relation between verbal content and one's emotional experience is rather complex. Nevertheless, with algorithms improving and automation made possible, body/face movement analysis and text-based analysis can become convenient choices.

Another promising measure is physiological sensors, which is a dominant approach in affective science and has been made more accessible outside conventional lab environments in recent years. Villanueva, et al. (2018), for example, collected close-to-real-time skin conductance data to approximate and compare the emotional experiences of engineering students across different design activities. Rieuf, et al. (2017) measured electrodermal activity to compare emotional arousal frequency of designers between an immersive design environment and a more traditional one. Psychophysiological research of emotion is more widely adopted in interaction design (Balters & Steinert, 2017, Paredes, et al., 2018) but is still nascent in the research of designers' emotions. Despite the convenience made by physiological access to emotion, we need more reflections around the question: *what is it that is being measured in relation to emotion?*. The lack of theoretical linkage would weaken the interpretation of analytical results. Broadly in design research, physiology is often used for advancing research of design cognition rather than emotion (Gero & Milovanovic, 2020).

Surprisingly, none of the design studies reviewed above examined the validity of their measures or applied multiple measures to increase validity. It must also be noted, as reviewed above, that different technological approaches favor different conceptual perspectives of emotion. For instance, facial expression analysis software and algorithms such as Affectiva (as used in Zhou, et al., 2019) and iMotions[1] is often used for classifying emotional states, which fits the emotion specificity view. In contrast, electrodermal activity-based arousal detection suits the emotion dimensionality view. The choice of measure and instrumentation often depends on the researcher's theoretical basis as well as practical considerations, such as availability of technological equipment and in-house expertise.

3. What are the research questions?

3.1. What theories are we building upon?

The current research aims to explore how to capture momentary emotions that are relevant and revealing of designers' experiences and processes in their respectively ecologically valid settings. We adopt theories of distributed cognition and situated action, in which situated knowledge, cognition and behavior are actively constructed in interaction with the changing socio-physical environment (Bamberger & Schön, 1983; Suchman, 1987; Hutchins, 1995; Kirsh, 2008; Heylighen & Nijs, 2014; Rieuf, et al., 2017; Lahlou, 2018). The view sees design thinking and doing as being constructed through physical, mental and social mediations. Bamberger and Schön (1983), in particular, highlights the "continuous mobility" of knowledge and knowing. Along the same line, others (Hutchins, 1995; Suchman, 1987; Lahlou, 2018) have demonstrated through case studies how workers' actions are augmented with, controlled by and channeled through both the situated and cultural-historical physical and social realities. This lens is contrasted with the traditional view, where knowledge is stable, high-level and decontextualized and to design is to enact a codified body of knowledge and skills. Although emotion is not explicitly integrated in the theories of situated cognition and action, there is no doubt that emotion is socioculturally constructed (Barrett, 2017). A situated view of emotion thus assumes social context plays a key role in the production and management of designer's momentary emotional responses, which in turn influence the evolving cognitive activity and social interaction.

3.2. What is the analytical lens?

Such a theoretical taking makes the measurement of situated emotion pointless if ecological validity is not prioritized. Moreover, we take the stance that there is not "a single, coherent and non-contradictory account of what happened", and that "real action is often ambiguous and may have multiple determinations" (Lahlou, 2018). Embracing ambiguity is not common analytical practice, yet we have to acknowledge that the analysis of emotional responses and actions may never be complete. However, references to the specific situation and designers' subjective experiences would lead to more complete interpretation and meaningful analysis (Xue & Desmet, 2019).

As a result, audio-visual materials of the specific contexts as well as designers' subjective experiences need to be recorded in addition to unobtrusive objective measures of emotion to enable repeated examination in context. The inductive, complexity-embracing nature of such a study would allow meaningful context-relevant connections to be made about designers' emotion between the otherwise disconnected physical, sociocultural, emotional and cognitive parameters.

Measuring emotional responses in the wild can be made possible by a few technological advances. Amongst the emotion measures reviewed above, speech acoustics and electrodermal responses

are the less obtrusive methods that are potentially practical in field studies. Other desirable methods, such as facial expression-based emotion analysis, would require consistent recording of a single participant's face at a perpendicular angle, which is difficult to realize in a real-world situation. But even with the feasibility of voice and sweat gland, their instrumentations in the wild would induce a lot of real-world noises which may make speech data difficult to clean up and electrodermal responses hard to interpret. In addition, the mapping between the basic physiological processes of emotion and experiences of emotion (i.e., what is felt) is still poorly understood (Barrett, et al., 2007). In order to complement and interpret the objective measures as well as to match our theoretical position, traditional field study methods, such as observation and interview, are considered imperative in the current study.

3.3 What is the research question?

Such a multimodal dataset of the work of experienced designers in the real world would allow us to inductively find *a good measure, in terms of validity, of the situated emotion of designers working in an ecologically valid setting.* More specifically, our primary research questions are: How to capture designers' situated emotion in natural design work environments? What and how well do the proposed mixed-methods approach assess designers' situated emotion?

The second question declares the need to address the how and why of the (lack of) concordance across experiential (retrospective self-report), physiological (speech prosody, EDA), and behavioral (video analysis) measures. In cases where subjective reports disagree with objective psychophysiological measures, the discrepancies are analyzed to understand to which extent the measures are still valid and whether meaningful implications can be drawn about designer behavior.

By seeking a good measure, inherently we are demonstrating why it matters. Would the current measures, in combination, reveal design behaviors that would otherwise be unseen, undervalued or misunderstood? We will analyze how the introduced measures capture high-arousal emotional experiences in particular, and how analyzing these experiences in a situated manner assists a deeper understanding of designer behavior. Perhaps also pertinent to the importance of capturing emotion in design research is the question of how emotional experiences, as approximated by these emotion measures, interfere with designers' emergent design action. We will discuss how designers' emotional expectations and displays guide design action and reveal about designers' beliefs of design expertise.

4. What is the study design?

An ideal dataset would have multimodal data of designers situated in real-world projects and natural work scenarios. However, while video/audio recording and skin conductance wearables are less obtrusive to design work at hand, they are often considered intrusive to corporate confidentiality. The difficulty of

video-recording design work in the real world due to confidentiality and other inconveniences has been overcome elsewhere (Christensen, et al., 2017).

4.1. Participants

Experienced designers were recruited based on accomplishments. They all started as engineers and went through rigorous human-centered design programs (from 1970s to 2019) from an Engineering department at a U.S. university. The designers were randomly paired to work on the project for three hours. None of them collaborated on any design project before. We will focus on three design dyads, because of space limitation. Table 1 shows the six designers' profiles. We used previously established design process conception (Morozov, et al., 2008; Atman, et al., 1999) to assess their confidence, frequency of engagement in each design step, as well as how supportive their work environment is towards each step of design work (Morozov, et al., 2008; Atman, et al., 2010). The designers have on average 20.3 years of design work experience. Despite large variance of years of design experience, they are on average highly confident (avg: 85.69 out of 100 with std: 4.95), receive positive work context support and engage frequently with different kinds of design work. See Appendix for detailed self-evaluation of the designers.

Designer[1]	**Gender**	**Design confidence**[2] (0-100)		**Supportive work context**[2] (0-100)		**Years of Exp**[3]	**Work engagement**	**Self-identified role**
		mean	SD	mean	SD			
Accom	Male	81.57	2.37	75.57	9.27	27	*Prototype, Communicate*: frequently; Other design steps: occasionally	Product Designer, Researcher
Diver	Male	80	0	80	0	40	*Define, Gather, Ideate, Communicate*: a few times a day; Other design steps: occasionally	Design Innovator, Product Designer, Consultant, Educator, Researcher, Magician
Analyte	Male	75.71	9.76	69.29	9.32	6	Every design step: a few times a day	Engineering Designer, Researcher, Research Scientist
Narra	Male	95.71	7.87	100	0	40	Every design step: frequently	Designer, Innovator, Designer, Consultant, Educator
Model	Female	86.43	3.78	87.86	3.93	4	*Prototype, Communicate*: a few times a day; *Ideate, Evaluate*: frequently; Other design steps: occasionally	Engineer, Designer, Innovator, Product Manager, Consultant, Educator, Research Scientist
Criti	Female	94.71	5.91	70.29	34.08	5	*Define, Gather, Ideate, Communicate*: a few times a day; Other design steps: occasionally	Engineer, Designer, Innovator, Consultant, Research Scientist, CEO
Average	N/A	85.69	4.95	80.50	9.4	20.3	N/A	N/A

1. Anonymous fake names are used to protect participant privacy
2. Results are averaged across all design steps.
3. Counting years of design work experience starts after graduation

Table 1. Designer Demographics and Design Experience

4.2. Study design workflow

In this study, we recreated a close-to-real-world project experience by having participants work on an ill-defined problem outside the lab, in the field. Given a set of materials, the participants are asked to create

solutions to radically improve the dining experiences of families with small children[2]. Materials contain pre-collected videos of eating scenarios of several families, an instruction sheet, a piece of news about working moms, an article about cultural differences, and a map of the design studio's surroundings. (see Appendix for the instruction sheet).

To realize "in the field" experience, we crafted a temporary design studio situated at a place with easy access to children's facility, kitchen and residences of families with small children. The designers were physically unconstrained from the design studio. Figure 1a and 1b shows examples from the study. Two hours into the study, two families with their children came in for user testing. We suggested the users not make compliments only, but provide honest, critical and helpful feedback to the designers. By the end of the three hours, each team delivered a pitch as if to their client with some prototypes in hand.

Figure 1a. Two designers working on improving dining experiences of families with small children in the design studio. It is a spacious children's playroom, with a large table, chairsets and whiteboard. The room is further equipped with a laptop-connected TV, various prototyping materials and tools. Outside the glass door is a home-style kitchen and a common area with sofa, tablesets, and TV. The participants are physically unconstrained and can go outside at any time during the study. The design studio is 1-minute walking distance from the residences of families with small children and a children playground.

Figure 1b. A designer sketching ideas about a water-refill function of the refrigerator for kids, viewed from the first-person perspective of the collaborating designer. This dyad is brainstorming in the kitchen.

4.3. Instrumentation and Data Collection

As the study prioritizes ecological validity, we employ multiple observational channels to account for uncontrolled variables, using first-person perspective cameras, fixed multi-angle cameras, audio recorder,

physiological sensor, and retrospective self-report. Fixed video cameras are located at the four corners of the room. Each participant wears a miniature eye-level video camera, so as to capture how situations are lived from the perspective of the participant (Lahlou, 2011; Lahlou, 2018). In addition, first-person perspective cameras would keep track of designers' activity as they step out of the room, freeing the intrusiveness of a cameraperson, as in the case in Figure 1b. In this study, we used AXON Flex 2[3] for accessing subjective views.

AXON Flex 2 also has a dual-channel audio that captures voice from a fixed position relative to the wearer's head. With that, we collected each participant's audio data for speech acoustic analysis. In particular, vocal intensity analysis is sensitive to audio recorder positioning and is made possible by the device's fixed position relative to the participant.

Each participant also wears an Empatica E4 wristband[4] on the dominant hand. The device acquires data of electrodermal activity[5], movement, skin temperature, heart rate variability, all with time reference.

Figure 2a. A design prototype as an example material used in the retrospective interview

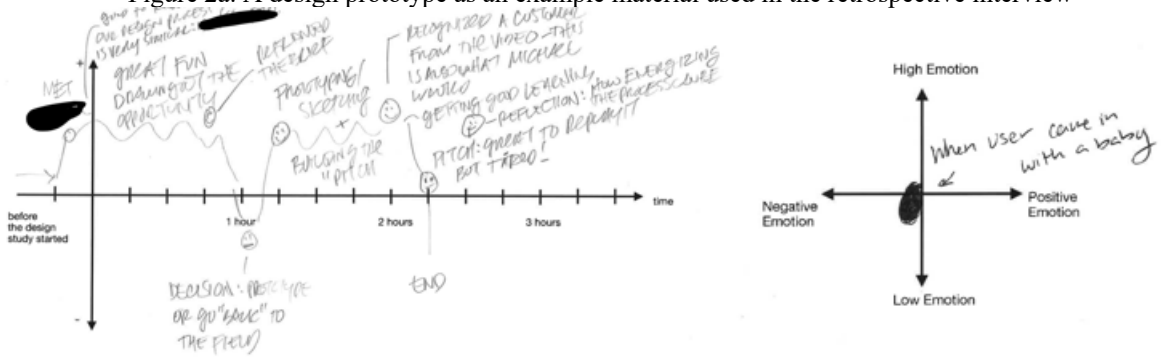


Figure 2b. Drawings of emotional valence map (left) and emotional valence and arousal matrix (right) as an example material used in the retrospective interview

Last but not least, participants shared subject experiences retrospectively in a 2-hour one-on-one interview. Large prints of snapshots of the participants' own views, artifacts produced during the project, subjective video clips and other tools were used when appropriate to facilitate the unfolding of internal experiences, as exemplified in Figure 2a and Figure 2b.

4.4 Data Analysis

Different measures of emotion are first cleaned up and analyzed separately, and then put together for comparative analysis and triangulation for micro-level analysis of salient events as well as high-level understanding of intrapersonal change and interpersonal dynamics.

Vocal pitch: Audio data for each participant is extracted from the headphone, manually cleaned up for speaker diarization and partitioned into utterances with ELAN software[6]. Laughing, whistling, sneezing and other sounds are separated from vocal speech, because they sometimes have very different pitch and intensity characteristics. With the cleaned-up audio data, vocal pitch and vocal intensity are extracted with PRAAT software[7]. It is noted that only voiced speech is used for vocal pitch (Detrich, 2014). The standard pitch range (30 to 450 for male, and 75 to 600 for female) is adopted except for some male participants who had rather high pitch and some other male participants have lots of creaky voices, where vocal pitch would be as low as 10 Hertz, and interfere with Praat's performance. To deal with the problems, we broadened the pitch range for these individuals, which does not affect data output (Detrich, 2014) but can eliminate the problem. We manually checked creaky voices with the Covarep program in Matlab software (Degottex, et al., 2014). To allow for between-subject comparison, the baseline was acquired by re-scaling vocal pitch to standard deviations above and below each designer's average vocal pitch (Detrich, et al., 2014). Utterances that are 1.5 standard deviations above the local average are regarded as pointers to pronounced episodes of surprise, confusion and curiosity. In addition, team-level speech data is also analyzed to understand individual behavior nested in dyads. Grid analysis (Brinberg, et al., 2018) is also applied to make explorative intergroup comparison from the perspective of one single emotional measure, such as vocal pitch.

Electrodermal activity: The EDA data retrieved from Empatica E4 (sampled at 4 Hz) was processed with Continuous Decomposition Analysis (CDA) as implemented in Ledalab with Matlab software. Skin Conductance Response (SCR) is exported using a minimum amplitude criterion of 0.05 μS. The minimum amplitude criterion is a threshold that the rise in skin conductance value must reach or surpass to qualify as SCR, typically set between 0.01 μS and 0.05 μS, depending on the recording condition. Since the current study took place in an uncontrolled environment, a score of 0.05 μS is used (Boucsein, 2012; Caruelle, et al., 2019). We use Skin Conductance Response (SCR) frequency to categorize the data into low, medium, and high levels (Boucsein, 2012; Dawson et al., 2017). More specifically, we use the standard that a frequency of 1–3 peaks/min (ppm) occurs at rest (Dawson et al., 2017, p. 225), and as frequency increases with the arousal level, values higher than 20 ppm are interpreted as high arousal (Boucsein, 2012, p. 222), while anything in between is labelled as medium. Signals less than 1 ppm are discarded for the purpose of data cleanup.

Video data and transcripts: The multi-angle video data is used for contextualization and triangulation with the above objective measures, and is analyzed openly without any specific coding system. Additionally, transcripts with reference to time are analyzed by the emotional lexicon collected, manually curated and evaluated by the National Research Council of Canada (NRC) (Mohammad & Turney, 2013). Because of our interest in emotional dimensionality rather than specificity, we used NRC lexicon to render results of valence, arousal and dominance of each utterance made by all the participants.

Retrospective self-report: The self-reports are transcribed first, and then coded using a specific lens (Saldaña, 2015) to retrieve self-perception of emotional experience along the design journey. The three coding categories are valence (positive - negative), arousal (high - low) and emotional states (self-reported by designers). For instance, in the process of scoping down the problem space, one designer said it was very "painful" — "there's an anxiety that you're wasting time… there's just like a little bit of fear involved in that, or anxiety". "It wasn't out of my comfort zone", the designer made it clear, that the anxiety thing "I've done a bunch of times", "I know what it feels… I know there are times you have to feel failure consciously like that. It's impending, and you just have to accept it… keep moving and then another sparkle comes and brings it back up, and that's something I'm familiar with". As a result, this synthesis phase is coded — valence: negative, arousal: mid-level. Not all participants were as articulative about their emotional experiences as this designer was. Nevertheless, most designers were able to talk about pronounced experiences where they experienced notable rises of stress or energy. All participants shared and evaluated their emotional experiences, with visual illustrations, provided thoughts regarding how and why they felt in a certain way. Phases where emotions were untold are marked as "unclear". Designers' emotion maps are used to assist matching the timeline, and case-based emotion matrices are used to mark self-rated valence and arousal more accurately. It is worth noting that the coding effort in the current study requires little interpretation, analysis and synthesis, as compared to inductive, qualitative studies. We did not do comparative coding across designers. In addition to emotion coding, self-reports are also used to retrieve the designers' general beliefs about emotion, behavior and ability both within and outside the current design task.

To test for concordance across different measures, we sliced the time-series data into meaningful chunks of design phases and social situations, when applicable, and compared the subjective score with the mean and variance of objective measures in each data segment using correlation and mean difference analysis. On the other hand, thinner slices of objective data allow for comparison about event-based observations with subjective reports. Where emotional experiences that are signified by participants (such as a big emotional change), analysis is done to see whether concordance is stronger. (additional note to review). Additionally, high-arousal moments as reflected by both pitch and SCR frequency measures are comparatively analyzed with video and subject report.

5. Results

5.1. Brief summary of the designers' retrospectively self-reported emotional experiences

The first dyad Narra and Analyte together designed three concepts. They spent the majority of the three hours on ideation and preparation for user-testing. In retrospect, Narra did not feel much changes of his emotion, except for two distinctive moments. In the first one, Narra was really angry at a, what he perceived to be a parenting problem, while watching the video materials. The second pronounced moment occurred when the parents, in the user-testing phase, disapproved of two prototypes from the team. One was a food-on-demand prototype designed for the kids to eat on their own schedules. The other was a meditation prototype designed to bring parents' attention to the dinner table and to their kids. In addition, Narra revealed that he struggled about breaking away from his previous experiences both as a seasoned children product designer and as a parent with a lot of opinions. For Narra's partner Analyte, he was less concerned about bringing in his own design assumptions. Analyte regarded the design work as "professional" where he applied "the process" and "rules of engagement". Analyte admitted he had a "oh, no" reaction as he first learned about the project, since he had no affinity for kids. He was neither a parent. Nevertheless, Analyte reported that he was slightly positive throughout. The only very emotional experience for him was during user testing. Because all their three concepts were designed for older toddlers, Analyte instantly felt "Darn! We didn't think of that" at the sight of one of the parents carrying a 1-year-old baby. Analyte was not comfortable with the unexpected situation in the user testing.

Different from Narra and Analyte who focused on solutions, The second dyad Diver and Accom spent most of their time defining the problem — analyzing and comparing user behaviors in video materials, relating to their previous experiences and drawing frameworks to make sense of the problem space. Also unlike the previous dyad, they did not explicitly plan their time, but rather sailed each other on the wind of their thoughts. Nevertheless, the team had an explicit design goal — to increase joy as a family. They invested a lot of time identifying needs lying in the gulf between "how to eat food with" (e.g., kids utensils) and "what to eat" (e.g., Plum space food). Diver reflected that he cared less about coming up with specific ideas, but rather identifying the right problem and generating the right need statement. Accom, however, wanted to have a few ideas to test with users. As a result, Diver and Accom had a conflict later in the process, which Diver described as "the high point of perplexity" and Accom regarded as the only negative experience he had during the process. In retrospect, Diver reflected how important it was to "keep the energy up". For Diver, the splitting point with Accom was a deflation of energy he gained from earlier explorations, and the project ended in "anticlimax" partly because of that.

Designer Criti and Model were yet another idiosyncratic design team. They were the only argumentative group that constantly talked over each other to disagree. In retrospect, they each reported

experiencing hiccups as they learned how to work together. Yet, the more they worked, the more highly they thought of each other. Indeed, their disagreements, though initially uncomfortable, seemed to give each other energy. A salient one for Criti occured in the middle of the project. She and Model had divergent directions moving forward, in terms of reassessing opportunity areas or pursuing the idea of "let the kids cook with a real knife". Criti insisted on pursuing the latter but at the same time felt lost because of Model's resistance. Criti retrospectively thought it was a "positive shift" that Model had suggested going to the kitchen to "feel out" some more ideas, which she was not convinced in the first place. Yet it effectively broke the undesirable equilibrium and re-energized the team. As for Model, the brainstorming phase at the kitchen was also a highlight for her. Model's passion was in materializing concepts, through sketching and building. Her other emotional highlight was at the prototyping phase, where she was able to engage with materials and tools to materialize one of her favorite ideas. Compared with the previous groups, this group balanced time amongst different design stages.

5.2. Identifying "oh no", "oh phew", and "oh yay" in situ with physiological measures

How well does the mixed method capture situated emotions-in-action of designers in the wild? We begin with the unique advantage of vocal pitch to identify design episodes of high engagements. Consider Diver's case. During the 3-hour-long design task, Diver had 1388 utterances (avg pitch: 157.59 Hertz). Using the 1.5-standard-deviation criteria, we identify 66 high-pitch utterances (avg: 276.37 Hertz). In the descriptions below, we use mean_std to denote how many standard deviations are above average vocal pitch. Subsequently, these utterances are identified in the multi-angle video materials, SCR frequency profile as well as retrospective self-reports for triangulatory analysis.

The analysis allows categorization of the high-arousal spikes of experiences into a few categories and are positioned in time as shown in Figure 3 - topleft. Out of the 66 high-voiced utterances, 9 are identified as expecting to end a situation or enter a new social situation, such as announcing "*Okay*!" (1.56 mean_std) to suggest moving onto the next design stage, and raising his tone to say "*Bye, thank you!*" (2.10 mean_std) to the parents at the end of the user testing.

The majority of the high-pitch talks has to do with social interaction — 23 are identified[8] as emotional interaction. For instance, an emotional team-bonding interaction occurred earlier in the design process (Figure 3 - Ex1) when Diver and Accom crystalized their ideas using a design tool both designers were very familiar with. Diver was delighted that Accom shared similar design training with him and joked, "*Okay, checked! Extra credit! We are done, we are out of here!*" (1.83 mean_std). The high-pitch utterance is followed by both designers in laughter. Diver described his delight he had every time he learnt a bit more about Accom as a "oh phew!" experience, because he was really worried about the

temporary teamwork with somebody he did not know before, and now learning that "there's a bunch of stuff both in language and process that I know we probably share" greatly relieved his concerns.

In addition to emotional interactions, clusters of emotional learning experiences are also identified. Emotional learning is suggested by fleeting moments of surprise, confusion and interest that spur new thoughts and actions, which are also surfaced by the vocal pitch measure and subsequently validated and understood through triangulation. We elaborate on two examples below.

> Figure 3 - Ex2 desirable surprise and confusion is a 2-minute episode of situated surprises and confusion for Diver as the team reviewed a video material of a child interacting with food and utensils. After watching the video in silence for a while, Diver eagerly shared his observation with Accom that adult products were scaled down for kids in softer materials; "*but it isn't clear*" (1.89 mean_std), he paused, and did not finish the sentence before continuing, "we've seen these modifications in adult world of plates and utensils. So within the culture we are trying to kidify it enough that (unfinished)... It isn't clear it's helping the kids. And I think the parents go, this looks like it is for kids". He continued with a related observation from past experiences that some adult bottle designs were actually influenced by baby products and that "it literally went the other way around!" As the video played along, he turned his attention back and forth between the TV screen and his partner. "*But I'm realizing it's a baby bottle!*" (1.53 mean_std), he excitedly pointed to the screen; "But that…", Diver quickly brought his hand back to hold his chin, and then stretched out to gesture, contradicting his earlier comment: "*and it works for her! It works perfectly. She had no problem with that!*" (1.66 mean_std). While Accom acknowledged Diver's observation, Diver kept watching the video, holding a fist by his chin, asking in a deeper and lower voice: "So are we looking at this the wrong way, that we wanted to have...", he briefly paused, turned toAccom, and increased his volume, taking a lighter tone, "kids become adults versus learn from the kids and modify our adult world (with laughter)! That might not be hard to sell! That might not be hard to sell."

Within this short time, Diver was channeled to speak in abnormally high voices a few times. He identified kidification as an interesting phenomenon as he watched the child interacting with utensils.

His body language suggests excitement; his high voice and repetitive verbal expressions (e.g., "*it isn't clear*") suggest confusion; his contradictory statements (e.g., "it isn't clear it's helping the kids" versus "it works for her!") suggest surprise and confusion; and his questioning — "are we looking at this the wrong way" — suggests strong curiosity. Instead of simply acknowledging kidification, Diver made it a rather intriguing issue. This physical material-elicited emotional learning experience is also evident from the perspective of SCR frequency, where a steep rise occurs in the local proximity of the video-watching activity, as shown in Figure 3 - bottomleft. This emotional episode of surprise, confusion and curiosity marks a time when Diver shifts his attention from parents' needs to kids' needs by questioning the existing baby products around "how to eat food with".

> Figure 3: Ex3 undesirable surprise and tension is a longer episode where Diver had a tense disagreement with Accom on what to do next after the team finished the video-based user-behavior observations. Diver suggested talking to actual users to validate their current design assumptions before moving onto ideation and prototyping. Diver had indicated a few times of his strong interest to talk with real users from the beginning on, and now he made it an explicit

> request. But instead of supporting Diver's suggestions as Accom always did until this point, Accom proposed to design some sacrificial concerts first. Hearing that, a high-pitch voice leaked from Diver before Accom finished his sentence. "Hmm! Um-hum!" (1.68 mean_std). Diver opened up to go with Accom's thinking process to ideate. After a few minutes of exploring the solution space, Diver was channeled back to think of talking to users, because one of their concepts was hinged on the question "when kids transition from baby food (to normal meals)". "That", points Diver to the TV screen, "looks like my plate, only it's not working!... And we literally had to separate ourselves from the kids because it doesn't work…Ah, I… I would love to ask about when you eat together, when you don't". In response, Accom insisted on spending the limited time on coming up with some concepts and preparing for user-testing. Diver attempted to persuade Accom again: "We've thrown some ideas out around this stuff, and I'd love to see — are we even close to be on track", and he added, "our hypothesis is that they [parents and kids] are struggling eating together. I just, I want to validate that". But Accom still persisted in his thinking. "*Okay*" (2.14 mean_std), responded Diver disappointedly.

In the first high-pitch reaction, Diver is verbally positive, but his vocal pitch suggests surprise and discomfort. As an experienced designer, Diver is able to regulate his emotions and make conscious balance between maintaining teamwork and making the right pathway. As Diver reflected in the retrospective interview, he regards the division with Accom as "the high point of perplexity". In Diver's view, the project ends in "anticlimax" partly because of that, even though he perceived that they received positive and encouraging feedback from the users. Diver was deeply convinced, based on his experience, that it was not wise to go too deep in concept generation without validating their underlying assumptions with actual users first. Intriguingly, Diver started navigating this "felt difficulty" with a positive attitude and open mind. During this process, there is no doubt that Diver was guided to reshape his understanding about the design project at hand, as new questions surfaced (e.g., when do kids transition from baby food to family food?). By cooperating on ideation, Diver built a new mental container that would accommodate product ideas while not stretching his own convictions too much. But that only led to building up more tension. At the end of the brief ideation when Accom expressed strong interest to work on concepts, Diver realized they have different conceptions around "concept" and "ideation" — while for Diver they already have some high-level ideas, to Accom these were not concrete concepts ready to be tested with users. This experience also has a pivotal effect on Diver's subsequent performance. He was less curious about the task and less engaged behaviorally as well. Diver laughed much less (Figure 3 - topleft) and showed fewer emotional learning behaviors as indicated by the high-pitch criteria.

Shifting lens to EDA, Diver's arousal does not make obvious sense as in Ex2, since phase 24 and 25 are actually lower than the local proximity, despite the high baseline. This could be better understood through contextualization. In Diver's reflection, he described design progress as "going along where your energy is" and emphasized how important it was to "keep the energy up". For Diver, the splitting point with Accom was a deflation of energy he gained from earlier activities. This is observable in his physical activation as well. Diver changed from the active posture of standing up and working by the whiteboard at

phase 22 and 23 back to sitting down, as he tried to resolve the conflict. In comparison to Ex2 in which Diver's high-pitch utterances suggest desirable surprise and curiosity, the high-pitch utterances of Ex3 are briefer in length and suggest undesirable surprise and tension.

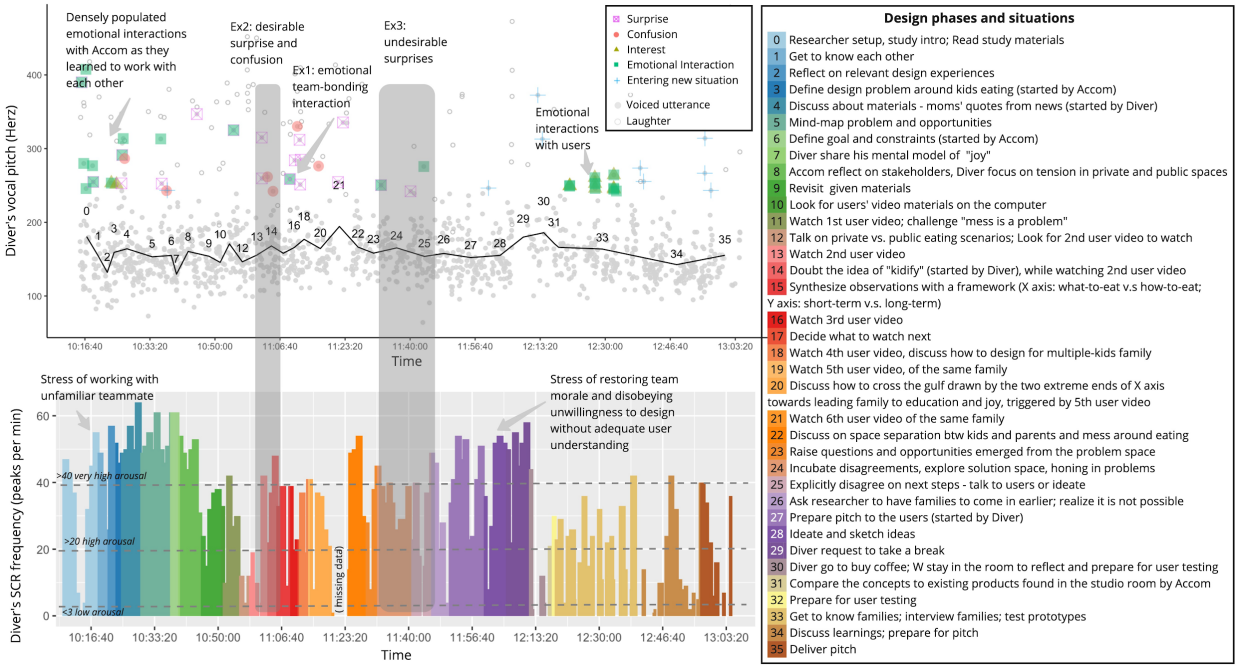


Figure 3. Diver's engagement map segmented by design phase and social situation; Topleft: average vocal pitch (error bars not shown for the sake of simplicity), Bottomleft: Skin conductance response (SCR) frequency, where determination of arousal level is based on standards (Boucsein, 2012; Dawson et al., 2017). The identification of high-emotional moments and episodes (as indicated by colored shapes in top-left graph) is based on 1.5 standard deviations above the designer's average vocal pitch, and subsequently qualitatively coded through triangulation with SCR, retrospective self-report and video analysis. Several examples of Diver's emotion experiences are highlighted and elaborated in the main body.

5.3. Intraindividual change, and inter-individual and inter-group difference

In addition to reliably pointing to designers' highly emotional engagements, bodily singlas of arousal level also track designers' emotional changes across time, and allow between-designer and between-group comparison.

5.3.1 Designers' emotional fluctuation covary with design phase

Intraindividual change can be characterized by patterns of emotional engagement within a designer that emerge from the mapping of physiological signals over time. A consistent pattern we find is the covariation of physiological arousals with design phases and/or social situations. Table 2 presents the case of designer Diver. More evidence can be found in the Appendix.

Design Stage	Design phase and situation*	Physiological measures	Subjective self-report and/or video analysis
Early exploration	phase 1 - 7	**Highest stress:** ❑ Densely populated with 40 peaks of SCR per min or more; ❑ Dense high-voice talks with Accom that are characterized as emotional interaction.	Diver regarded the early interaction with the unfamiliar partner Accom as most stressful, because "figuring out the partner was the most important piece of getting through [the design project]" and he was really worried they did not have a "shared unspoken process".
Observation and needfinding	phase 11 - 21	**Highest surprise and confusion:** ❑ Densely-populated high-pitch verbal expressions that are coded as surprise and confusion.	Diver was constantly stimulated by user behaviors in engagement with the video materials
Ideation and preparation for user-testing	phase 27 - 29	**Highest stress:** ❑ Densely populated with 40 peaks of SCR per min or more.	This stage is immediately after the tense disagreement, which Diver regarded as "the high point of perplexity". Here, Diver was under great stress of restoring team morale by disobeying his unwillingness to design products without a full understanding of user needs.
User testing	phase 32 - 33	**High interest:** ❑ Densely-populated high-pitch verbal expressions that are coded as interest.	Diver sharpened his tone a few times to ask the parents questions, e.g., "*Do you guys ever eat together at that table or do you eat separately?*" (mean_std: 1.67). Diver reflected that the answer to this question was very critical to validating his assumption of the product idea.
Reflection and Pitch delivery	phase 34 - 35	**Lowest engagement** ❑ Lowest SCR frequency and lowest average vocal pitch	Diver believed the time after needfinding was not spent correctly. As a result, the project ended in an "anticlimax", as reflected by Diver.

* phase numbering refers to Figure 3.

Table 2. Designer Diver's emotional fluctuation covary with design phase.

Let us elaborate on the ideation stage as an example for illustrating Table 2 and showing how situated emotional experiences were identified and validated. Diver's SCR frequency peaked in phases 27 to 29, the ideation stage, which is immediately after the tense disagreements between the designers. From the behavioral perspective, the stage started with Diver walking across the table to sit side-by-side with Accom and joining Accom to sketch ideas. And at the same time Diver said, "I won't leave you, [laughter], no no no" (Figure 4). Thereafter, Diver and Accom behaviorally were quite engaged in ideation and refining concepts for the user-testing. However, as Diver revealed in the interview afterwards, he firmly believed it was not wise to generate concepts without validating their underlying assumptions with actual users first. Still, he chose to do so because team morale was a higher priority for him. While Diver regulated his outwardly teamwork behaviors, such as responding positively to Accom (e.g., "Oh, I like that!") and normalizing tones, his slowly-developed stress was revealed in the SCR frequency graph. As shown in Figure 3 - bottomleft, this stage is densely populated with 40 peaks of SCR per min or more. Unlike the kind of stress of working with an unfamiliar partner in the exploration stage, this time, Diver was under great stress of restoring team morale by disobeying his unwillingness to design products without a full understanding of user needs. Right afterwards, Diver went outside for a coffee

break (phase 30), a point in time that his emotion transitioned to a much lower arousal, as shown in the SCR frequency graph in Figure 3.

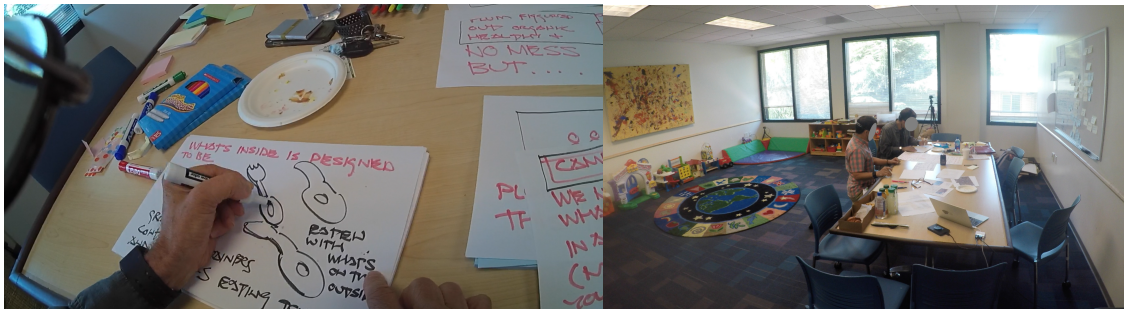


Figure 4. After the explicit disagreement on what to do next - whether to talk to users or to come up with concepts, Diver broke the line by coming to the other side of the table and joining Accom in ideation. As he walked over, he said "I won't leave you, [laughter], no no no". This was the first time Diver and Accom sat so close to each other. They were engaged in idea generation and preparation for user testing thereafter. Left: first-person perspective of Diver; right: third-person perspective.

5.3.2 Interpersonal difference

Dyad	Designer	Num of utterance	Vocal Pitch Mean (HZ)	Vocal Pitch SD (HZ)	Vocal Intensity Mean (dB)	Vocal Intensity SD (dB)	SCR Frequency Mean (peaks/min)	SCR Frequency SD (peaks/min)
1	Narra	1635	118.29	30.54	57.71	11.49	25.48	13.94
1	Analyte	1492	117.52	35.79	59.49	12.28	6.85	7.68
2	Diver	1388	157.59	54.73	60.93	12.81	30.37	17.12
2	Accom	917	103.81	35.02	58.8	13.88	13.69	10.88
3	Model	1646	255.03	84.03	65.25	11.99	23.67	14.51
3	Criti	1630	273.3	84.76	65.94	12.61	28.77	15.07
	Average	1451.3	170.92	54.15	61.35	12.51	21.47	13.2

Table 3. Descriptive results of vocal pitch and SCR frequency

Comparison of emotional behaviors between designers allows us to revisit a particular designer to better understand his/her behavior. Table 3 gives the descriptive results of vocal pitch and SCR frequency. Note that vocal pitch is not standardized and thus is not meant for comparison. For instance, female designers Model and Criti had higher pitch mean than other male designers, which should reflect more of a gender difference than arousal difference. Based on Table 3, we can see Accom spoke much fewer in terms of number of utterances than other designers.

Below, we focus on a few cases to exemplify interpersonal comparison. Figure 5 gives an example comparing Diver and Analyte. Opposite to Diver, B's arousal stays low to medium throughout the design task except for only a few salient phases. Admittedly, individual difference plays a role in EDA measure, in that Analyte might have been less sensitive to EDA measure. On the other hand, the differences between Diver's high arousal and B's low arousal can be better understood by analyzing their retrospective reflection. Diver cared about keeping the energy up by immersing himself into the work, and he described design progress as "going along where your energy is". In comparison, Analyte tended to detach himself from the design work, as for him, the project was foremost a "professional" one, and it

was all about applying "the process" and "rules of engagement". The drastically different framings and and design intentions allow us to make sense of the contrasting arousal levels of the two designers. Notably, Diver's arousal in the user-testing phase is low relative to his average SCR frequency. The dipping matches with Diver's subjective experience of energy deflation, as analyzed above. In comparison, Analyte's SCR frequency peaked in the same phase, where he had a "oh no" experience as his rules of engagement were broken by an unexpected user-testing situation, as we report earlier in section 5.1.

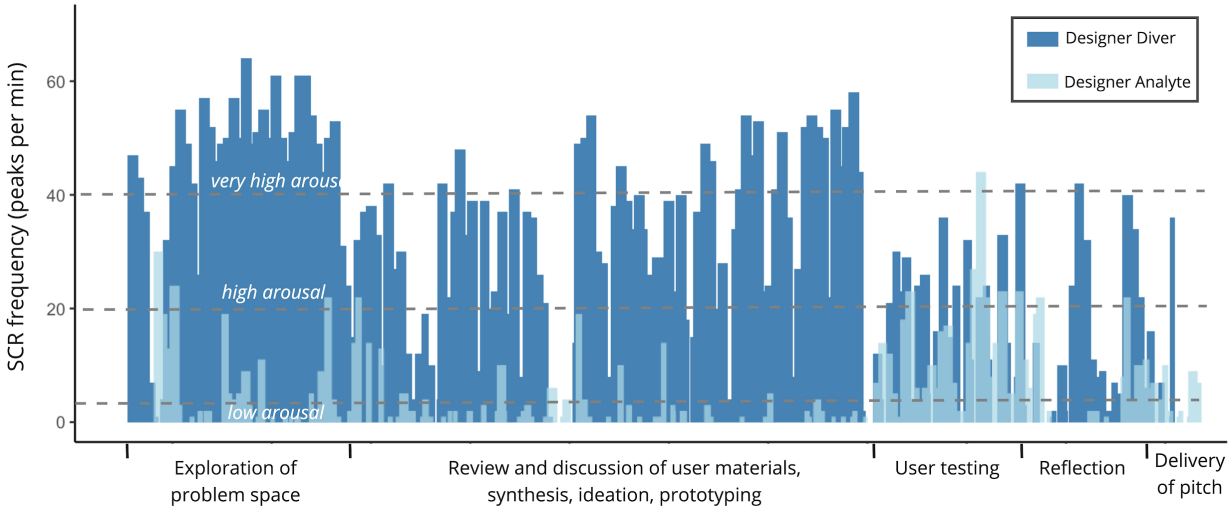


Figure 5. Comparison of SCR frequency-based emotional arousal between Diver and Analyte who have distinctively different arousal profiles. Notably, Diver's arousal in the user-testing phase is low, whereas in the same phase it is highest for Analyte.

This between-subject pattern in SCR frequency is not visible in vocal pitch, as shown in Figure 6. Most of the design talks are well regulated and normalized, as shown in the grey points of utterances that sediment towards the means, making it a poor tool to sense gradual changes in emotional arousal. But even with its insensitivity, the pitch difference between Diver and Analyte in the user-testing phase is significant ($p = 0.000$, one-tailed, unpaired t-test, where Analyte had 181 valid utterances, avg: 0.39, and Diver had 143 valid utterances, avg: 0.11). In fact, Diver had the lowest pitch in average amongst the two teams. Whereas all the other three designers raised their average vocal pitch as they interacted with the users, Diver dropped his average vocal pitch from previous stages. Here, designers' vocal pitches are standardized to be standard deviations from mean, to allow for between-subject comparison.

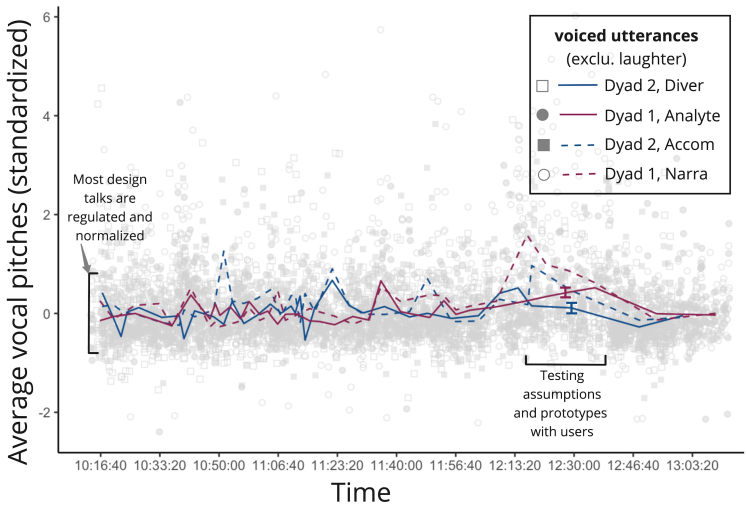
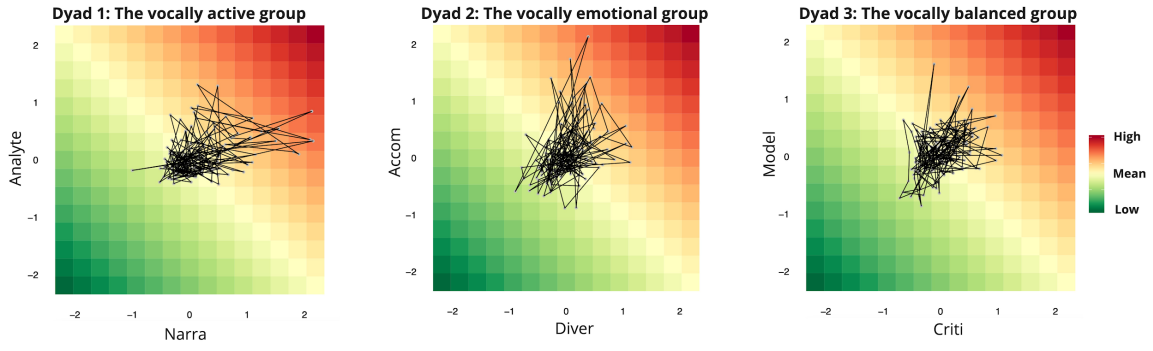


Figure 6. Standardized vocal pitch of designers from dyad 1 and 2 who shared the same work time. Most of the design talks are fairly well regulated, as shown in the gravitating grey points. Most designers spoke in a relatively higher tone in the user testing phase, when both teams received users' critical feedback. The exception is Diver. Difference between Analyte and Diver is highlighted in the graph.

5.3.3. Intergroup comparison

In addition to intraindividual and interpersonal analysis, team-level analysis allows for high-level comparison of designers' emotional experience. Characterizing team-level engagement is important for understanding each designer's individual emotional experience, because every designer's emotional behavior is partly shaped by the close social interaction with their teammates. In Table 3, the mean and standard deviation of SCR frequency shows both dyad 1 and dyad 2 were unbalanced so that one designer is in general much more stressed or energetic than the other, whereas the two designers in dyad 3 were more synchronized. Figure 7 shows how grid analysis of vocal pitch helps with making intergroup comparison from the perspective of vocal pitch.



Note: x and y axes represent standardized vocal pitch

Figure 7. Emotion dynamics heatmap. Here, each dyad-level emotional arousal is a time-series dataset of vocal pitch, each data point is an average pitch of utterances minute by minute. The graph shows how emotional dynamics is different across groups. Dyad 1 can be characterized as the vocally active team, for it had relatively fewer arousals below 0, and its activeness is largely driven by designer Narra. Dyad 2 and 3 are more balanced in terms of high and low arousals. Dyad 2 is more spread out than dyad 3, with more high pitches. This indicates dyad 2 may have experienced more emotional ups and downs than dyad 3.

5.4. Do different emotion measures converge?

We have shown how the mixed-methods approach assesses designers' situated emotions. But how good is it? We will focus on addressing the validity challenge below.

The straightforward validity check approach is to examine how well different measures synchronize within each designer. Our statistical results confirm that different objective and subjective measures of emotion tend not to match with each other in general. More specifically, one-tailed, unpaired t-test analysis suggests concordance between SCR frequency and vocal pitch in terms of difference between high and low arousal is observed in two of the designers, Criti and Analyte (for both, $p<0.001$), but not found in other designers. In the matching cases of Criti and Analyte, rising vocal pitch in a certain design phase would match with rising SCR frequency of the same design phase.

Concordance between NRC-based emotion analysis in texts and vocal pitch is found in four of the designers, but not in others. The four designers are Criti (pitch correlates with negative emotional word frequency per utterance: $p = 0.052$), Model (pitch correlates with high-arousal verbal content: $p<0.001$ and positive expressions: $p<0.001$), Analyte (pitch correlates with positive expression: $p=0.006$), and Diver (pitch correlates with high-arousal verbal contents: $p=0.037$, and with both positive expression: $p=0.005$ and negative expression: $0=0.038$). In addition, phase-by-phase correlation between SCR frequency and vocal pitch is only found in Analyte ($r=0.62$, $p=0.014$).

We also get mixed results about the concordance between retrospective reports and measures of SCR frequency and vocal pitch. More specifically, we compare different measures based on a few pronounced experiences as reported by the designers in retrospect. In the first dyad, Narra reported a sense of failure when testing the prototypes, especially the first one, with users, with which one may interpret the spiking of vocal pitch in the user-testing phase as a sign of Narra's surprise and discomfort at users' reactions. Narra's SCR frequency graph has a significant jump during the 1st prototype-testing phase. The pitch measure also caught Narra's reportedly angry experience, but this particular experience is in the shadow of the unvaryingly high SCR frequencies early in the design process. In comparison, Analyte's highly emotional moment was consistently revealed in both vocal pitch and EDA, making Analyte the only designer whose multimodal measures are synchronized across board.

For the second group, Accom's vocal pitch profile reveals momentary surprises as he interacted with the project material and families in the user-testing stage. Interestingly, neither vocal pitch nor EDA shows strong signs of Accom's reportedly difficult experience of dealing with the tense conflict with Diver. In comparison, Diver had a few abnormally high-pitch utterances as he dealt with the conflict, but otherwise, Diver's average vocal pitch did not fluctuate much by design phase either. Similar to Narra's failure experience in user testing, Diver also had a dip of SCR frequency, which curiously matches with Diver's subjective experience that it was a deflation of energy.

For the last design group, Model's EDA measure recorded the highest SCR frequencies during the prototyping phase, which confirms her subjective experience. Model's other pronounced emotional experiences, such as brainstorming in the kitchen and receiving feedback from users, were also visible in the EDA measure. But none of her subjectively pronounced experiences were distinguishable in the vocal pitch graph. For Criti, the SCR frequency profile suggested she was highly aroused most of the time, and it captured a few occasions when Criti's arousal plummeted, including when Criti felt lost in the middle of the process. But this emotional experience of Criti's seems to be normalized in verbal expressions. What is made salient from Criti's vocal pitch graph is a heated interaction with Model, which contributed to Criti's positive impression about Model.

In total, 73.3% of the most pronounced emotional experiences, as reported by the designers, find validation in the respective SCR frequency graphs, and 40% of them co-occur with rising or dropping pitch average. All comparative analysis and physio-measure graphic illustrations can be found in the Appendix.

5.5. Making sense of situated emotion using conditional concordance

We have shown that different physiological and expressive emotion measures do not always concord within each designer. In order to make sense of the discordance, the peculiarities of different measures are analyzed. We find SCR frequency reliably reflects a designer's perceived level of energy and stress, which is often otherwise regulated in vocal expressions. For example, Narra is shown to have more than 20 peaks of SCR per min most of the time, his high energy manifested by his physical activeness. Yet, Narra's vocal pitch analysis displays a drastically different pattern suggesting he was verbally regulated most of the time.

In comparison, vocal pitch best captures involuntarily leaked momentary surprises, confusions and interests. Where spikes of experiences are surfaced by vocal pitch, however, the average vocal pitch of utterances in the corresponding design phase is not necessarily high, because again these experienced designers regulated and normalized most of their design talks. For instance, a notably high-pitch episode of Diver occurred in design phase 14, which we elaborated in section 5.2., but the average vocal pitch of phase 14 does not differ significantly from other phases. On the other hand, comparison of vocal pitch averages of different design phases and/or social situations does shed light on how designers' emotions fluctuate across stage and situation. Phase-based vocal pitch average is especially telling of situational change, which channels the designer to talk differently. This is evident in either significantly higher vocal pitch or larger variance as the designer shifts from work to taking a break, takes a pause to pose questions to the researcher, or transitions from brainstorming to prototyping.

In addition, the physiological data also reveals emotional behavior. We have shown that Accom and Analyte were the only two designers low in energy and stress level except for salient phases during the design project. This may reflect designers' different ways of emotional engagement in professional settings in general. We have observed outside the study that Diver and Criti, for instance, also present outwardly active and energetic behaviors in other design activities. Analyte, in contrast, confirmed his not-easily-aroused temper in the post-task interview.

Lastly, in terms of verbal content, NRC-based text analysis of emotional valence and arousal does not display obvious patterns across individuals, but shows strong idiosyncratic patterns within some designers. It is interesting, for instance, high-voiced talks are reliably characterized by positive word usage for Analyte, whereas designer Criti uses more negative words when she raises her voice. The seemingly inconsistency, however, explains why Analyte and Criti respectively had highly emotional experiences in each of their situations — Analyte spoke positively as he proactively dealt with his surprise and stress during the user-testing phase, whereas Criti spoke negatively as she made critical comments on parenting behaviors during the needfinding stage.

In sum, when integrated with observational data and self-reports, the three objective measures of emotion offer complementary views for researchers to see and understand designers' emotional experiences more comprehensively, as demonstrated so far. The analysis of discordance amongst the measures, in particular, provides an opportunity to ask new questions and see the data differently, and therefore demonstrates the value of conditional concordance. We thus propose the notion of conditional concordance and define it as achieving coordination and convergence amongst different, sometimes discordant, emotional measures through triangulation and contextualization.

We have offered a few examples of conditional concordance in the sections above. In order to make the point clearer, consider the mismatch in Narra's case between his subjective experience and the EDA measure at the needfinding stage. The drop of SCR frequency at the time when Narra reportedly felt really angry at the parenting problem urges us to ask why. The seemingly contradiction, however, is understandable once we dig deeper into Narra's experience across time. Similar to Diver, Narra was observed to be physically active in the beginning. His excitement and stress were attributed by himself to working with a much younger partner he never worked with before. He was burdened with the thought that Analyte would be uncomfortable working with himself, as he was "30 years older" and much more experienced. As a result, Narra had a high baseline of SCR frequency, and the parenting problem that made him really angry was buried in the high baseline. However, the stimulus was easily caught in vocal pitch, which is a reliable sensor of involuntarily leaked high emotional response. Because this user-behavior observation aroused a relatively minor emotional reaction in comparison to teamwork-related concern, one can reasonably conjecture that Narra did not accumulate more stress as a result of watching

the parenting problem. Nevertheless, this user-behavior observation did drive Narra to design educational products for parents, as manifested by one of dyad 1's prototypes.

6. Discussion

So far, we have shown in a few steps how to identify and characterize momentary emotional responses and situated emotional episodes by the proposed mix-methods approach. In this section, we discuss the emotion measures under the larger context of affective science, reflect on the study design, and draw implications for design research and design practice.

6.1. How is emotion measured again?

As shown in Figure 8, vocal pitch and skin conductance response are objective, real-time, highly-grained, context-independent measures of one emotional dimension (e.g., arousal); In contrast, retrospective self-report is subjective, inconcurrent, event-contingent, prone to inconsistent granularity and lacks sharp time boundaries. In contrast to objective measure, retrospective self-report gives access to designers' internal experiences and would capture the sociocultural complexity of emotion with greater symbolic value. To resolve the incomparability challenge, we choose to compare objective and subjective measures based on the two to five most memorable and remarkable emotional experiences as reported by the designers. The qualitative data also offers critical contextual information to make sense of the series of lifeless numbers that represent vocal pitch and EDA.

Designers characteristically used laughters, whistles, coughing, creaky voices, and singing during their design projects. Due to their different acoustic behaviors, we did not consider them as valid voiced utterances for vocal pitch analysis. In addition, overlapping talks were not included because of the chosen analytical approach. Apart from pitch, vocal intensity, talking time, silence, turn-taking, and other vocal acoustics are also important but not analyzed in the paper.

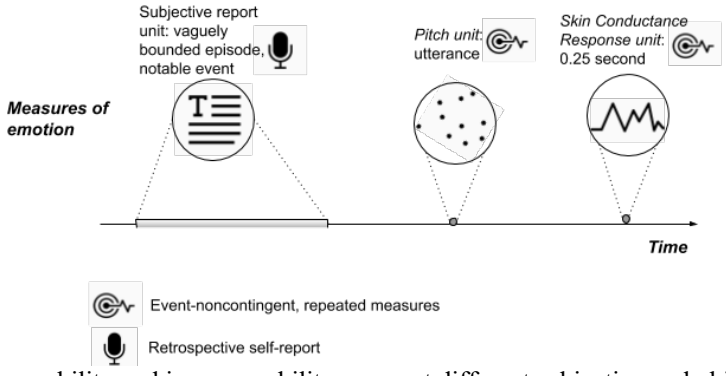


Figure 8. The comparability and incomparability amongst different subjective and objective measures that directly operated on participants' emotion. The interview-based self-report, without careful nudging from the researcher to have participants talk about details, tends to focus on a high-level set of recollections, such as "I didn't have any negative emotions throughout", or "the prototyping part was pretty good", which does not directly translate into the level of detail of objective data.

Why is it important to have multiple investigative lenses? The under-defined construct emotion is composed of complex social, cultural, psychological and biological elements. We have applied some instrumentation and analytical tools from affective science. In affective science, however, scholars tend to avoid applying different measures due to poor concordance (Barrett, 2017). The inconsistent result of emotion concordance in our study is consistent with findings in previous affective science research. Essentially, any one measure of emotion has its own limitations and biases and is likely associated with variances unique to it (Mauss & Robinson, 2009). However, as a result of choosing a single measure, construct validity of arousal or stress is often unaddressed in affective science.

We argue that the validity challenge should not become the reason for not pursuing multivariate methods, because it can be adequately addressed by a complexity-embracing methodological approach. We propose the notion of conditional concordance through contextualization and triangulation as a powerful tool to work with multimodal data. The current method is robust to measures that may be sensitive to real-world noises, such as EDA, and is thus suitable for research where ecological validity has nonnegotiable high priority. To this end, we have provided ample evidence how this complexity-embracing approach captures ecological momentary emotion with comprehensiveness and accuracy, and thus improves validity.

Validity will remain to be a challenge in emotion research. Emotion can sometimes have mixed and even contradicting components simultaneously (Barrett, 2017). This is reflected in the current study. For instance, Diver reported he was most stressful earlier in the process because of working with an unfamiliar partner. But at the same time, he was pretty neutral and relaxed about the project itself, which is obscured in the physiological measures. It is also curious that Accom's physiological measures were more revealing of his engagement with project materials and users, but fail to show adequate evidence of his reported difficulty working with Diver as their conflict incubated in the middle of the design task. This may be attributed to the fact that Accom felt at the same time glad and relieved, as he reported in the interview, that Diver compromised and the project went the way he wanted. How to capture the nuances and contradictions of emotional experiences could be an important future direction.

6.2. Reflections on the study design in the wild

Traditionally, lab studies prioritize measurable, standardizable and repeatable conditions over ecological validity. While lab studies make possible a wide range of psychological measures and special instrumentations, from computer-based assessment to fMRI-based neurological measures, the rigid codification techniques employed in lab studies may greatly distort the phenomenon researchers intend to study (Crilly, 2019; there are exceptions, see Sirkin, et al., 2015). Individuals are to a large extent

detached from the actual emotional stimulation the real-world context would invoke. In this regard, many of the studies above fall short of ecological validity (Ram, et al., 2017). Field study, at the opposite end of the ecological validity continuum, has the potential to address the issue. Partly due to the impracticality of many lab-based measures (e.g., facial behavior-based analysis) as well as potential problems caused by subjectivity and lack of standardization, field study is still an underrepresented method within the design research community (Lawson, 2004; Crilly, 2019, a few examples in p84 – p86).

The current research has been conceived to apply the obvious advantage of lab studies in ecological settings. Our research has a strong bias towards representing real-world design to reveal the mundane realities of design. In post-study reflective interviews, a few participants voluntarily expressed how real the experience was. In particular, the involvement of real users gives the subjects real-life pressure to work well and deliver well. In the meantime, the idea of participating in a research study dissolved gradually, in particular due to the in-artificiality of the environment, as reported by the participants. In analogy to automotive simulator-based studies, we have successfully created a design experience simulator. Designers are channeled to experience disturbances in such a study setting.

Admittedly, the setup of the study may not reflect a "normal" design project, especially due to the short project time (i.e., 3 hours). However, there might not be a normal project in a fast-changing world, and designers need to adjust their expectations and modify their behaviors to deal with the idiosyncrasies of every new project. Therefore, the ability of modifying behavior and attitude, but not necessarily knowledge, in accordance to the specific project is part of designers' ability.

6.4. Implications for design research and design practice

Historically, design researchers have had more access to and more emphasis on cognition and behavior of design. This bias is shared across disciplines. Lev Vygotsky (1962, cited in Roth & Lee, 2017) wrote more than five decades ago:

> We have in mind the relation between intellect and affect. Their separation as subjects of study is a major weakness of traditional psychology since it makes the thought process appear as an autonomous flow of 'thoughts thinking themselves,' segregated from the fullness of life, from the personal needs and interests, the inclinations and impulses, of the thinker. (p. 8)

The systemic bias has to be overcome, and emotion should no longer be treated as a byproduct of a designer's social-technical activity (Martin, Knopoff, and Beckman, 1998).

One of the biggest implications from our work is to see designers as emotional, growth-seeking individuals. The commonsensical but undermined idea of emotional ups and downs during the design process is strongly supported by statistical analysis of the time-series physiological data. Even for Diver and Accom who have 40 years of design experience, they did not just shuffle through, but experienced intense emotional ups and downs that directly affected their performance. In addition, as shown in the

current study, a designer situated emotional behavior is revealing of his/her knowledge and expertise. This is made evident, for instance, from Diver getting curiously stuck on a seemingly trivial problem, Analyte sticking to professional rule engagement, and Narra's emotional reaction to a parenting issue that eluded the eyes of most other designers. For design practitioners, we hope our research would materialize into a practical tool to raise emotion awareness. One might want to self-reflect what emotional behaviors are (not) conducive to desired project outcome or learning outcomes. We hope to facilitate fruitful conversations at the intersection of design thinking, feeling, doing and learning.

Together, we find vocal pitch a reliable pointer to surface pronounced momentary emotions as a result of dealing with new, unexpected stimuli. We show that a designer's emotional arousal, as measured by vocal pitch and SCR frequency, has meaningful differences across design phases as well as social situations. In addition to intraindividual change, the physiological measures also offer an analytical lens to understand interpersonal and intergroup differences. Text-based emotional measure, on the other hand, has the potential to reveal a designer's idiosyncratic verbal behaviors. We show that these objective emotional measures have unique advantages of their own and are complementary to each other as well as to traditional observational and self-report measures.

7. Conclusion

In a nutshell, we have presented a novel mixed-method approach and shown its unique investigative power of surfacing designers' situated emotion in uncontrolled, naturalistic settings. The inductive, complexity-embracing nature of the mixed-methods approach asks researchers to cast a wider net of investigative lenses, so that meaningful context-relevant connections can be made between the otherwise disconnected physical, sociocultural, emotional and cognitive parameters. Our research challenges how emotion should be studied within and outside design research and highlights the importance of emotion research for enhancing design practice. It would hopefully facilitate a fruitful decolonial conversation between design researchers and emotion researchers outside our field.

Endnote

1. https://imotions.com/biosensor/fea-facial-expression-analysis/
2. The prompt gives no constraints on what solution to deliver. Most design tasks in past research offer participants a defined solution scope, such as a new trash system (Dorst, 2015), a carrying/fastening device to fasten and carry a backpack on a mountain bicycle (Cross, et al., 1996; Cross & Cross, 1998), a playground (Atman, et al., 1999), or ask participants to redesign an object (Cannon, 2018).
3. https://www.axon.com/products/axon-flex-2
4. https://www.empatica.com/en-int/research/e4/
5. What does EDA measure? Electrodermal activity (EDA, previously known as galvanic skin response) refers to the variation of the electrical conductance of the skin in response to sweat

secretion, which is produced by the sympathetic nervous system. Skin momentarily becomes a better conductor of electricity when either external or internal stimuli occur that are physiologically arousing. Arousal is a broad term referring to overall activation and is widely considered to be one of the two main dimensions of an emotional response. For more, refer to Boucsein, 2012 and others.

6. https://archive.mpi.nl/tla/elan
7. https://www.fon.hum.uva.nl/praat/
8. Note that the categorization is not mutually exclusive. For instance, the new-situation example of saying bye to the users is also an example of emotional interaction.

Reference

Ahmed, S., & Wallace, K. M. (2004). Understanding the knowledge needs of novice designers in the aerospace industry. *Design studies*, 25(2), 155-173.

Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. *Administrative science quarterly*, 50(3), 367-403.

Anzaldúa, G. (1987). *Borderlands/la frontera* (Vol. 3). San Francisco: aunt lute books.

Atman, C. J., Chimka, J. R., Bursic, K. M., & Nachtmann, H. L. (1999). A comparison of freshman and senior engineering design processes. *Design studies*, 20(2), 131-152.

Balters, S., & Steinert, M. (2017). Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *Journal of Intelligent Manufacturing*, 28(7), 1585-1607.

Bamberger, J., & Schön, D. A. (1983). Learning as reflective conversation with materials: Notes from work in progress. *Art Education*, 36(2), 68-73.

Barrett, L. F. (1997). The relationships among momentary emotion experiences, personality descriptions, and retrospective ratings of emotion. *Personality and Social Psychology Bulletin*, 23(10), 1100-1110.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1), 1-23.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Bartel, C. A., & Saavedra, R. (2000). The collective construction of work group moods. *Administrative Science Quarterly*, *45*(2), 197-231.

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative science quarterly*, 47(4), 644-675.

Behoora, I., & Tucker, C. S. (2015). Machine learning classification of design team members' body language patterns for real time emotional state detection. *Design Studies*, 39, 100-127.

Bennett, C. L., Peil, B., & Rosner, D. K. (2019, June). Biographical Prototypes: Reimagining Recognition and Disability in Design. *In Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 35-47).

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4), 602-607.

Brinberg, M., Ram, N., Hülür, G., Brick, T. R., & Gerstorf, D. (2018). Analyzing dyadic data using grid-sequence analysis: Interdyad differences in intradyad dynamics. *The Journals of Gerontology: Series B*, *73*(1), 5-18.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of emotions*, *2*, 173-191.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1), 18-37

Cannon, D. M. (2018). *Prediction of Design Team Performance Through Analysis of Language Use in Meetings*. Stanford University.

Caruelle, D., Gustafsson, A., Shams, P., & Lervik-Olsen, L. (2019). The use of electrodermal activity (EDA) measurement to understand consumer emotions–A literature review and a call for action. *Journal of Business Research*, 104, 146-160.

Casakin, H., & Goldschmidt, G. (1999). Expertise and the use of visual analogy: implications for design education. *Design studies*, 20(2), 153-175.

Chai, C., Cen, F., Ruan, W., Yang, C., & Li, H. (2015). Behavioral analysis of analogical reasoning in design: Differences among designers with different expertise levels. *Design Studies*, 36, 3-30.

Chiu, M. L. (2003). Design moves in situated design with case-based reasoning. *Design studies*, 24(1), 1-25.

Christensen, C. M. (2013). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Christensen, B. T., Ball, L. J., & Halskov, K. (2017). *Analysing design thinking: Studies of cross-cultural co-creation*. CRC Press.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. *In Flow and the foundations of positive psychology* (pp. 35-54). Springer, Dordrecht.

Csikszentmihalyi, M. (2013). *Flow: The psychology of happiness*. Random House.

Clancey, W. J. (2016). *Creative engineering: Promoting innovation by thinking differently, by John E. Arnold.* Edited with an introduction and biographical essay by William J. Clancey.

Coyne, R. (2005). Wicked problems revisited. *Design studies*, 26(1), 5-17.

Crane, J. (2015). 'The bones tell a story the child is too young or too frightened to tell': The Battered Child Syndrome in Post-war Britain and America. *Social History of Medicine*, 28(4), 767-788.

Crilly, N. (2015). Fixation and creativity in concept development: The attitudes and practices of expert designers. *Design Studies*, 38, 54-91.

Crilly, N., & Firth, R. M. (2019). Creativity and fixation in the real world: Three case studies of invention, design and innovation. *Design Studies*, 64, 169-212.

Crocker, J., Fiske, S. T., & Taylor, S. E. (1984). Schematic bases of belief change. *In Attitudinal judgment* (pp. 197-226). Springer, New York, NY.

Cross, N. (2004). Expertise in design: an overview. *Design studies*, 25(5), 427-441.

Cross, N., & Cross, A. C. (1998). Expert designers. *In Designers* (pp. 71-84). Springer, London.

Cross, N., Dorst, K., & Christiaans, H. (Eds.). (1996). *Analysing design activity*. Wiley.

Cross, N. (2008). *Engineering Design Methods: Strategies for Product Design*.

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 1-36.

Damasio, A. R. (1994). *Descartes' error*. Penguin Group (USA), Inc.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2017). *The electrodermal system.* In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Cambridge handbooks in psychology. Handbook of psychophysiology* (p. 217–243). Cambridge University Press.

Davis, M. A. (2009). Understanding the relationship between mood and creativity: A meta-analysis. *Organizational behavior and human decision processes*, 108(1), 25-38.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014, May). COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 960-964). IEEE.

Dong, A., Kleinsmann, M., & Valkenburg, R. (2009). Affect-in-cognition through the language of appraisals. *Design Studies*, 30(2), 138-153.

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, 22(5), 425-437.

Dorst, K. (2015). *Frame innovation: Create new thinking by design*. MIT press.

Dym, C. L., & Little, P. (1999). *Engineering design: A project-based introduction*. John Wiley and sons.

Epstein, D. (2019). *Range: Why Generalists Triumph in a Specialized World*. Riverhead Books.

Eyal, G. (2019). *The crisis of expertise*. John Wiley & Sons.

Ewald, B., Menning, A., Nicolai, C., & Weinberg, U. (2019). Emotions Along the Design Thinking Process. In *Design Thinking Research* (pp. 41-60). Springer, Cham.

Fayard, A. L., Stigliani, I., & Bechky, B. A. (2017). How nascent occupations construct a mandate: The case of service designers' ethos. *Administrative Science Quarterly*, 62(2), 270-303.

Ge, X., & Leifer, L. (2020). When tough times make tough designers: how perplexing experiences shape engineers' knowledge and identity. *The International journal of engineering education*, 36(2), 650-663.

Gerber, E., & Carroll, M. (2012). The psychological experience of prototyping. *Design studies*, 33(1), 64-84.

Gero, J. S., & Milovanovic, J. (2020). A framework for studying design thinking through measuring designers' minds, bodies and brains. *Design Science*.

Gino, F., & Ariely, D. (2012). The dark side of creativity: original thinkers can be more dishonest. *Journal of personality and social psychology*, 102(3), 445.

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert?. *Intelligence*, 45, 34-45.

Hariharan, B. (2011). *Innovating Capability for (Deweyan) Continuity of Inquiry in the Face of (Zimbardoean) Discontinuity Within the Context of Engineering Education Research: Fostering Collaborations with Underserved Communities in the Developing Regions of the World*. Department of Mechanical Engineering Stanford University.

Harris, A. (1983). The intellectual standing of engineering design. *Design Studies*, 4(3), 147-150.

Heylighen, A., & Nijs, G. (2014). Designing in the absence of sight: Design cognition re-articulated. *Design Studies*, 35(2), 113-132.

Hu, Q., Bezawada, S., Gray, A., Tucker, C., & Brick, T. (2016). Exploring the link between task complexity and students' affective states during engineering laboratory activities. In *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers Digital Collection.

Hutchins, E. (1995). *Cognition in the Wild* (No. 1995). MIT press.

Hutchinson, A. (2018). *Designers, Emotions, And Ideas: How Graphic Designers Understand Their Emotional Experiences Around Ideation*. A Dissertation at Wayne State University

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, 24(4), 602-611.

Johri, A., Williams, C. B., & Pembridge, J. (2013). Creative collaboration: A case study of the role of computers in supporting representational and relational interaction in student engineering design teams. *International Journal of Engineering Education*, 29(1), 33-44.

Juhl, J., & Lindegaard, H. (2013). Representations and visual synthesis in engineering design. *Journal of Engineering Education*, 102(1), 20-50.

Jung, M. F. (2016). Coupling interactions and performance: Predicting team performance from thin slices of conflict. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(3), 1-32.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 263-286.

Kessous, L., Castellano, G., & Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, *3*(1-2), 33-48.

Kitayama, S. E., & Markus, H. R. E. (1994). *Emotion and culture: Empirical studies of mutual influence* (pp. xiii-385). American Psychological Association.

Kilker, J. (1999). Conflict on collaborative design teams: understanding the role of social identities. *IEEE Technology and Society Magazine*, 18(3), 12-21.

Kirsh, D. (2008). Problem Solving and Situated Cognition. *In P. Robbins & M. Aydede (Eds.), The Cambridge Handbook of Situated Cognition* (Cambridge Handbooks in Psychology, pp. 264-306). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511816826.015

Kunda, G. (2009). *Engineering culture: Control and commitment in a high-tech corporation.* Temple University Press.

Lahlou, S. (2018). *Installation theory: The societal construction and regulation of behaviour*. Cambridge University Press.

Lang, P. J. (1988). What are the data of emotion?. In *Cognitive perspectives on emotion and motivation* (pp. 173-191). Springer, Dordrecht.

Lawson, B. (2004). Schemata, gambits and precedent: some factors in design expertise. *Design studies*, 25(5), 443-457.

Lawson, B., & Dorst, K. (2013). *Design expertise*. Routledge.

Leifer, L., & Meinel, C. (2019). Looking further: design thinking beyond solution-fixation. *In Design Thinking Research* (pp. 1-12). Springer, Cham.

Lougheed, J. P., Brinberg, M., Ram, N., & Hollenstein, T. (2020). Emotion socialization as a dynamic process across emotion contexts. *Developmental psychology*, 56(3), 553.

Martin, J., Knopoff, K., & Beckman, C. (1998). An alternative to bureaucratic impersonality and emotional labor: Bounded emotionality at The Body Shop. *Administrative Science Quarterly*, 429-469.

Martinez, A. M. (2017). Visual perception of facial expressions of emotion. *Current opinion in psychology*, *17*, 27-33.

Mauss, I., Wilhelm, F., & Gross, J. (2004). Is there less to social anxiety than meets the eye? Emotion experience, expression, and bodily responding. *Cognition and Emotion,* 18(5), 631-642.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion*, *23*(2), 209-237.

Mesquita B & Kawasaka M. (2002). Different emotional lives. *Cognition and Emotion*, 16:127–41

Micheli, P., Wilner, S. J., Bhatti, S. H., Mura, M., & Beverland, M. B. (2019). Doing design thinking: Conceptual review, synthesis, and research agenda. *Journal of Product Innovation Management*, 36(2), 124-148.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436-465.

Morozov, A., Kilgore, D., Yasuhara, K., & Atman, C. (2008, June). Same courses, different outcomes? Variations in confidence, experience, and preparations in engineering design. In *Proceedings of the American Society for Engineering Education Annual Conference* (pp. 2008-768).

Moscovici, S. (1984). The Phenomenon of Social Representations, pp. 3-69 *in R.M. Farr and S. Moscovici (eds) Social Representations*. Cambridge, UK: Cambridge University Press.

Neeley, W. L. (2007). *Adaptive design expertise: A theory of design thinking and innovation* (Doctoral dissertation, Stanford University).

Op't Eynde, P., & Turner, J. E. (2006). Focusing on the complexity of emotion issues in academic learning: A dynamical component systems approach. *Educational Psychology Review*, 18(4), 361-376.

Paletz, S. B., Schunn, C. D., & Kim, K. H. (2011). Intragroup conflict under the microscope: Micro-conflicts in naturalistic team discussions. *Negotiation and Conflict Management Research*, 4(4), 314-351.

Paredes, P. E., Ordonez, F., Ju, W., & Landay, J. A. (2018, April). Fast & furious: detecting stress with a car steering wheel. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Paton, B., & Dorst, K. (2011). Briefing and reframing: A situated practice. *Design Studies*, 32(6), 573-587.

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology*, 106(3), 696.

Petriglieri, G., Ashford, S. J., & Wrzesniewski, A. (2019). Agony and ecstasy in the gig economy: Cultivating holding environments for precarious and personalized work identities. *Administrative Science Quarterly*, 64(1), 124-170.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98-125.

Purcell, A. T., & Gero, J. S. (1996). Design and other types of fixation. *Design studies*, 17(4), 363-383.

Pychyl, T. A., Lee, J. M., Thibodeau, R., & Blunt, A. (2000). Five days of emotion: An experience sampling study of undergraduate student procrastination. *Journal of social Behavior and personality*, 15(5), 239.

Ram, N., Gerstorf, D., Lindenberger, U., & Smith, J. (2011). Developmental change and intraindividual variability: Relating cognitive aging to cognitive plasticity, cardiovascular lability, and emotional diversity. *Psychology and aging*, 26(2), 363.

Reisenzein, R. (2000). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition & Emotion*, *14*(1), 1-38.

Rieuf, V., Bouchard, C., Meyrueis, V., & Omhover, J. F. (2017). Emotional activity in early immersive design: Sketches and moodboards in virtual reality. *Design Studies*, 48, 43-75.

Rahman, H. A., & Barley, S. R. (2017). Situated redesign in creative occupations–An ethnography of architects. *Academy of Management Discoveries*, 3(4), 404-424.

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Research in Human Development*, 14(3), 253-270.

Reimlinger, B., Lohmeyer, Q., Moryson, R., & Meboldt, M. (2019). A comparison of how novice and experienced design engineers benefit from design guidelines. *Design Studies,* 63, 204-223.

Roth, W. M., & Lee, Y. J. (2007). "Vygotsky's neglected legacy": Cultural-historical activity theory. *Review of educational research*, *77*(2), 186-232.

Runco, M. A. (1999). Tension, adaptability, and creativity. *Affect, creative experience, and psychological adjustment*, 165-194.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological bulletin*, *115*(1), 102.

Safin, S., Dorta, T., Pierini, D., Kinayoglu, G., & Lesage, A. (2016). Design Flow 2.0, assessing experience during ideation with increased granularity: A proposed method. *Design Studies*, 47, 23-46.

Sanger, S. (2012). *Breaking free: A qualitative analysis of entrenchment and disruptive strategies of corporate leaders* (Qualitative research report for the Doctor of management program).

Sas, C., & Zhang, C. (2010). Investigating emotions in creative design. *In Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design* (pp. 138-149). Desire Network.

Safin, S., Dorta, T., Pierini, D., Kinayoglu, G., & Lesage, A. (2016). Design Flow 2.0, assessing experience during ideation with increased granularity: A proposed method. *Design Studies*, 47, 23-46.

Shaw, M. P. (1994). Affective components of scientific creativity. *Creativity and affect*, 3-43.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Sirkin, D., Mok, B., Yang, S., & Ju, W. (2015). Mechanical ottoman: how robotic furniture offers and withdraws support. *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 11-18).

Smith, S. M., Linsey, J. S., & Kerne, A. (2011). Using evolved analogies to overcome creative design fixation. In Design creativity 2010 (pp. 35-39). *Springer*, London.

Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. *Cambridge university press*.

Todorova, G., Bear, J. B., & Weingart, L. R. (2014). Can conflict be energizing? A study of task conflict, positive emotions, and job satisfaction. *Journal of Applied Psychology*, *99*(3), 451.

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of personality and social psychology*, *90*(2), 288.

Ullen, F., Hambrick, D. Z., & Mosing, M. A. (2016). Rethinking expertise: A multifactorial gene–environment interaction model of expert performance. *Psychological bulletin*, 142(4), 427.

Valentine, M. A., Nembhard, I. M., & Edmondson, A. C. (2015). Measuring teamwork in health care settings: a review of survey instruments. *Medical care*, 53(4), e16-e30.

Vallet, F., Eynard, B., Millet, D., Mahut, S. G., Tyl, B., & Bertoluci, G. (2013). Using eco-design tools: An overview of experts' practices. *Design Studies*, 34(3), 345-377.

Vanasupa, L., Burton, R., Stolk, J., Zimmerman, J. B., Leifer, L. J., & Anastas, P. T. (2010). The systemic correlation between mental models and sustainable design: Implications for engineering educators. *International Journal of Engineering Education*, 26(2), 438.

Valkenburg, R., & Dorst, K. (1998). The reflective practice of design teams. *Design studies*, 19(3), 249-271.

Villanueva, I., Campbell, B. D., Raikes, A. C., Jones, S. H., & Putney, L. G. (2018). A multimodal exploration of engineering students' emotions and electrodermal activity in design activities. *Journal of Engineering Education*, 107(3), 414-441.

Voigt, R., Podesva, R. J., & Jurafsky, D. (2014). Speaker movement correlates with prosodic indicators of engagement. In *Speech prosody* (Vol. 7).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Xue, H., & Desmet, P. M. (2019). Researcher introspection for experience-driven design research. *Design Studies*, 63, 37-64.

Yilmaz, S., & Seifert, C. M. (2011). Creativity through design heuristics: A case study of expert product design. *Design Studies*, 32(4), 384-415.

Youmans, R. J., & Arciszewski, T. (2014). Design fixation: Classifications and modern methods of prevention. *AI EDAM*, 28(2), 129-137.

Zhou, J., Phadnis, V., & Olechowski, A. (2019, August). Analysis of Designer Emotions in Collaborative and Traditional Computer-Aided Design. *In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 59278, p. V007T06A043). American Society of Mechanical Engineers.

Appendix

1. An example showing the project workflow with matched personal reflection.

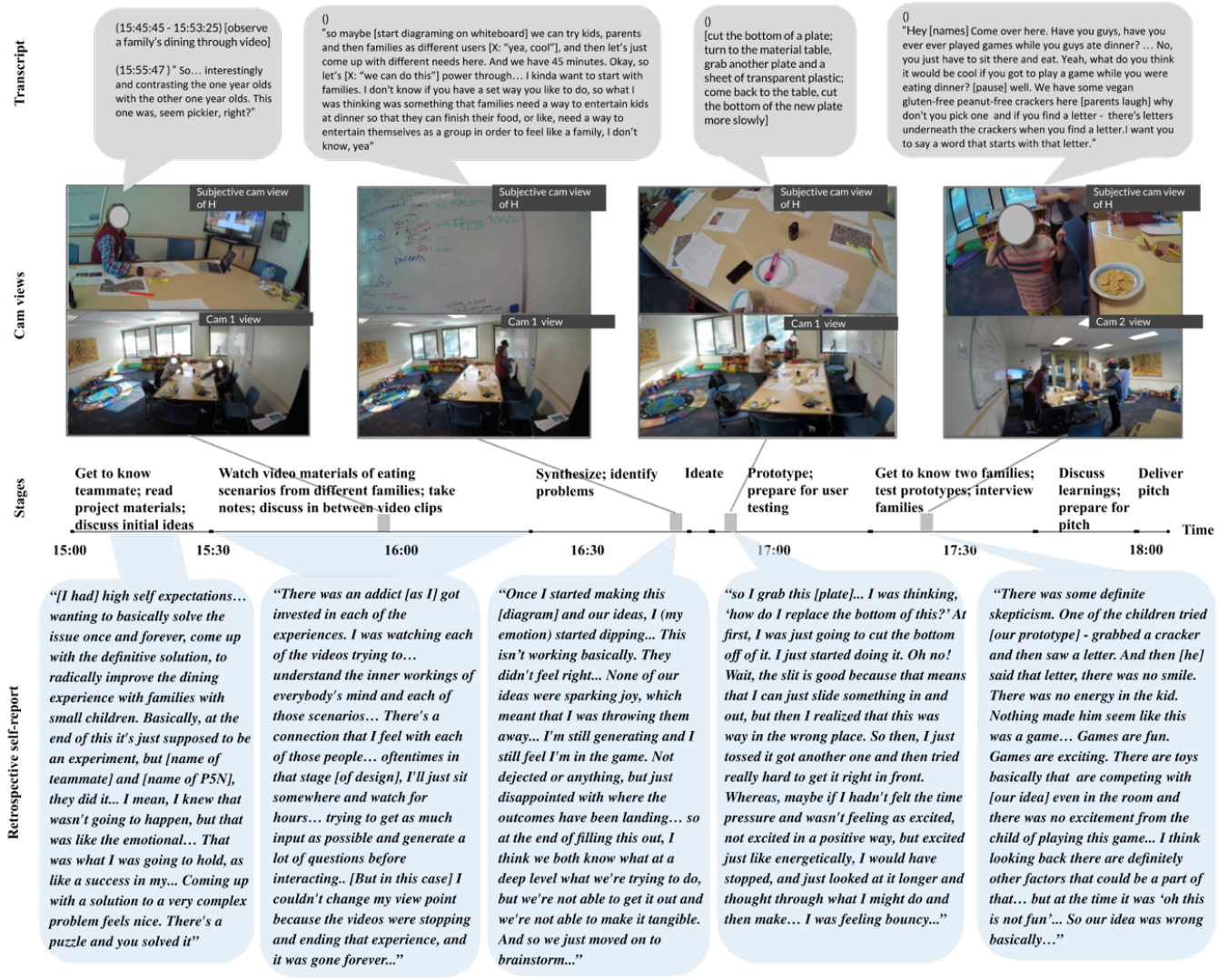


Figure 9. study work process (the texts need to be edited) and qualitative data

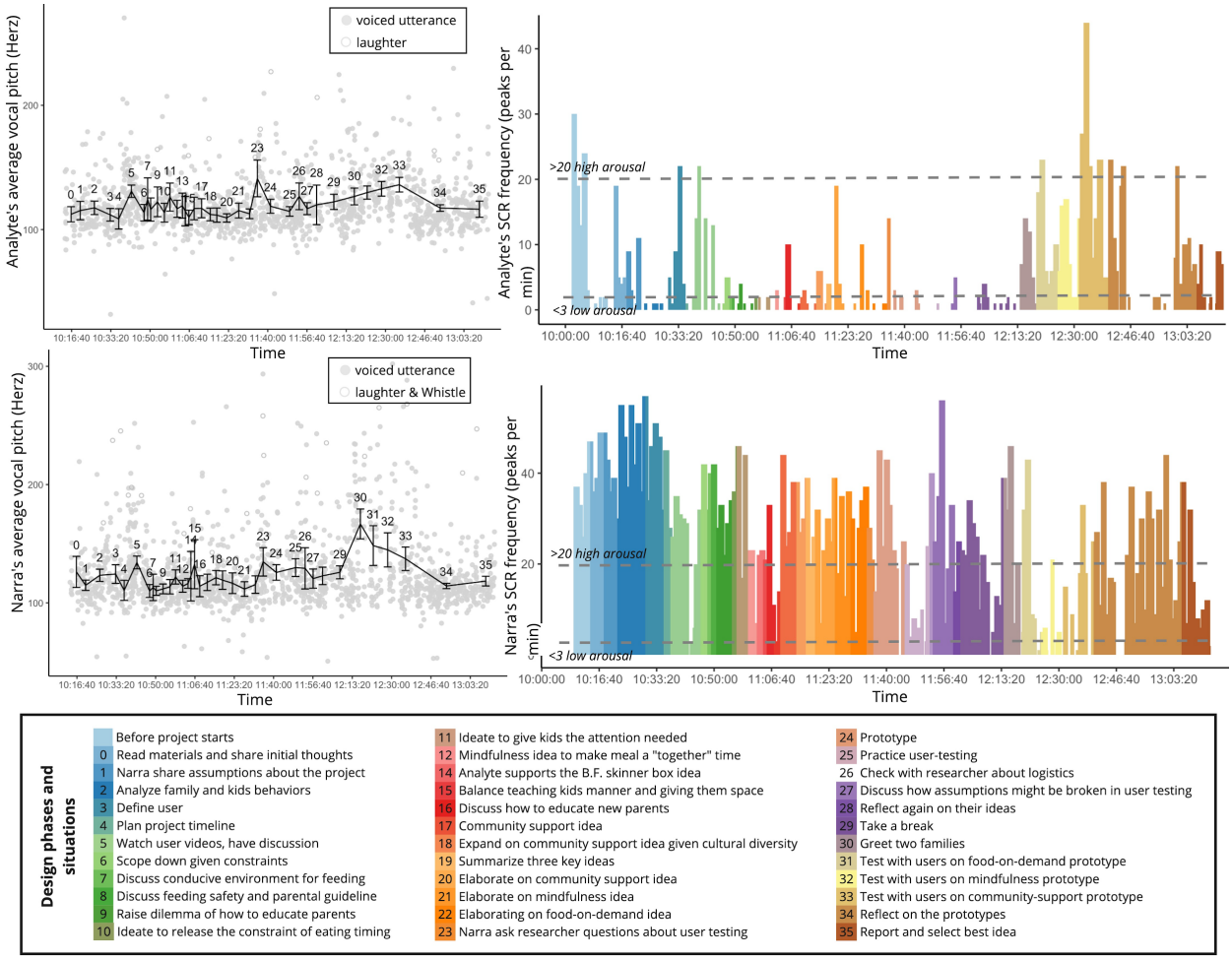
2. Example of emotional arousal and valence analysis based on retrospective self-report

Phase	Valence	Arousal	Accounts (emotional states highlighted)
Start	pos	high	Designer H commented a few times that at the start he was "very ***excited***", and that reading and thinking about the project "got me really **high**".
...			
Observation	neg	unclear	H's theories were disproved as he watched more user materials, yet no new sparkles arrived — "I don't feel so good", "very ***lost***".
Synthesis	neg	mid	In the process of scoping down the problem space and framing the need statement, H said it was "***painful***" — "there's an ***anxiety*** that you're wasting time… there's just like a little bit of ***fear*** involved in that, or anxiety". "It wasn't out of my comfort zone", he made it clear, that the anxiety thing "I've done a bunch of times", "I know what it feels… I know there are times you have to feel ***failure*** consciously like that. It's impending, and you just have to accept it… keep moving and then another sparkle comes and brings it back up, and that's something I'm familiar with".
Ideation	pos + neg	high	Before H and his teammate started brainstorming, they made the last try of coming up with potential need statements, "I was feeling at that point ***very energetic***, slightly positive, which was like ***anticipatory***, like ***hopeful*** and ***creative***".
	neg	mid	But "it isn't working" — the right verbiage wasn't emerging. That made him "dip down" to the first big negative point. He felt "***disappointed***", but he was "still above the halfway point in terms of energy" and felt he was "still generating".
Prototyping	pos + neg	high	Once they moved into prototyping, "I would say that energy was even higher, at its highest ever" of the whole journey. He explained that this was partly because of the time pressure — "I needed to make this thing before the families [come]". "I was just moving very fast, and I was constantly looking for things..." And he pointed in terms of valence, his emotion was "maybe just back and forth right around" the midpoint. He felt "***bouncy***", trying out how to model/prototype the idea, he felt generative, and as he was getting "my own personal feedback on the idea", he started to question it, "which brought me down negatively a little bit".
...			

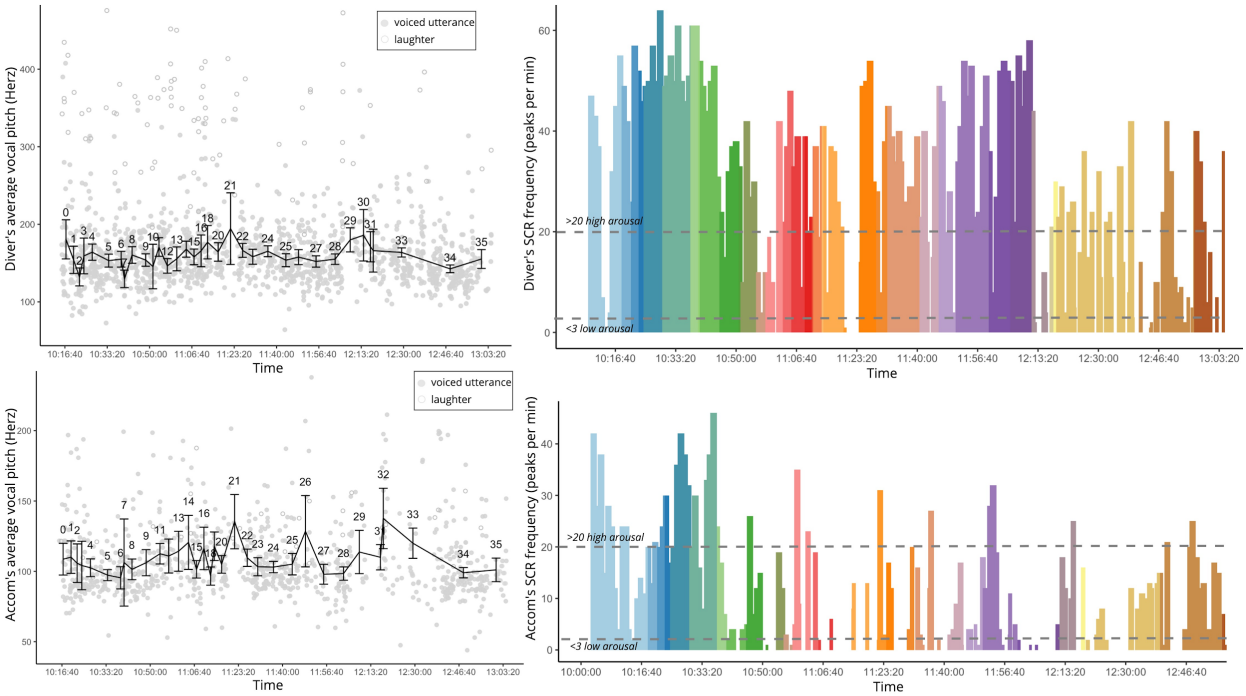
Table 4. Coding emotional responses from retrospective self-report based on one designer. Not all participants were as articulative about their emotional experiences as this designer was. Nevertheless, all participants shared and evaluated their emotional experiences, with visual illustrations, provided thoughts regarding how and why they felt in a certain way. Phases where emotions were untold are marked as "unclear".

3. Vocal pitch and SCR freq graphs of designers of the three introduced dyads in this paper.

Dyad 1



Dyad 2



Design phases and situations

0 Researcher setup, study intro; Read study materials
1 Get to know each other
2 Reflect on relevant design experiences
3 Define design problem around kids eating (started by Accom)
4 Discuss about materials - moms' quotes from news (started by Diver)
5 Mind-map problem and opportunities
6 Define goal and constraints (started by Accom)
7 Diver share his mental model of "joy"
8 Accom reflect on stakeholders, Diver focus on tension in private and public spaces
9 Revisit given materials
10 Look for users' video materials on the computer
11 Watch 1st user video; challenge "mess is a problem"
12 Talk on private vs. public eating scenarios; Look for 2nd user video
13 Watch 2nd user video
14 Doubt the idea of "kidify" (started by Diver), while watching 2nd user video
15 Synthesize observations with a framework (X axis: what-to-eat v.s how-to-eat; Y axis: short-term v.s. long-term)
16 Watch 3rd user video
17 Decide what to watch next
18 Watch 4th user video, discuss how to design for multiple-kids family
19 Watch 5th user video, of the same family
20 Discuss how to cross the gulf drawn by the two extreme ends of X axis towards leading family to education and joy, triggered by 5th user video
21 Watch 6th user video of the same family
22 Discuss on space separation btw kids and parents and mess around eating
23 Raise questions and opportunities emerged from the problem space
24 Incubate disagreements, explore solution space, honing in problems
25 Explicitly disagree on next steps - talk to users or ideate
26 Ask researcher to have families to come in earlier; realize it is not possible
27 Prepare pitch to the users (started by Diver)
28 Ideate and sketch ideas
29 Diver request to take a break
30 Diver go to buy coffee; W stay in the room to reflect and prepare for user testing
31 Compare the concepts to existing products found in the studio room by Accom
32 Prepare for user testing
33 Get to know families; interview families; test prototypes
34 Discuss learnings; prepare for pitch
35 Deliver pitch

Dyad 3

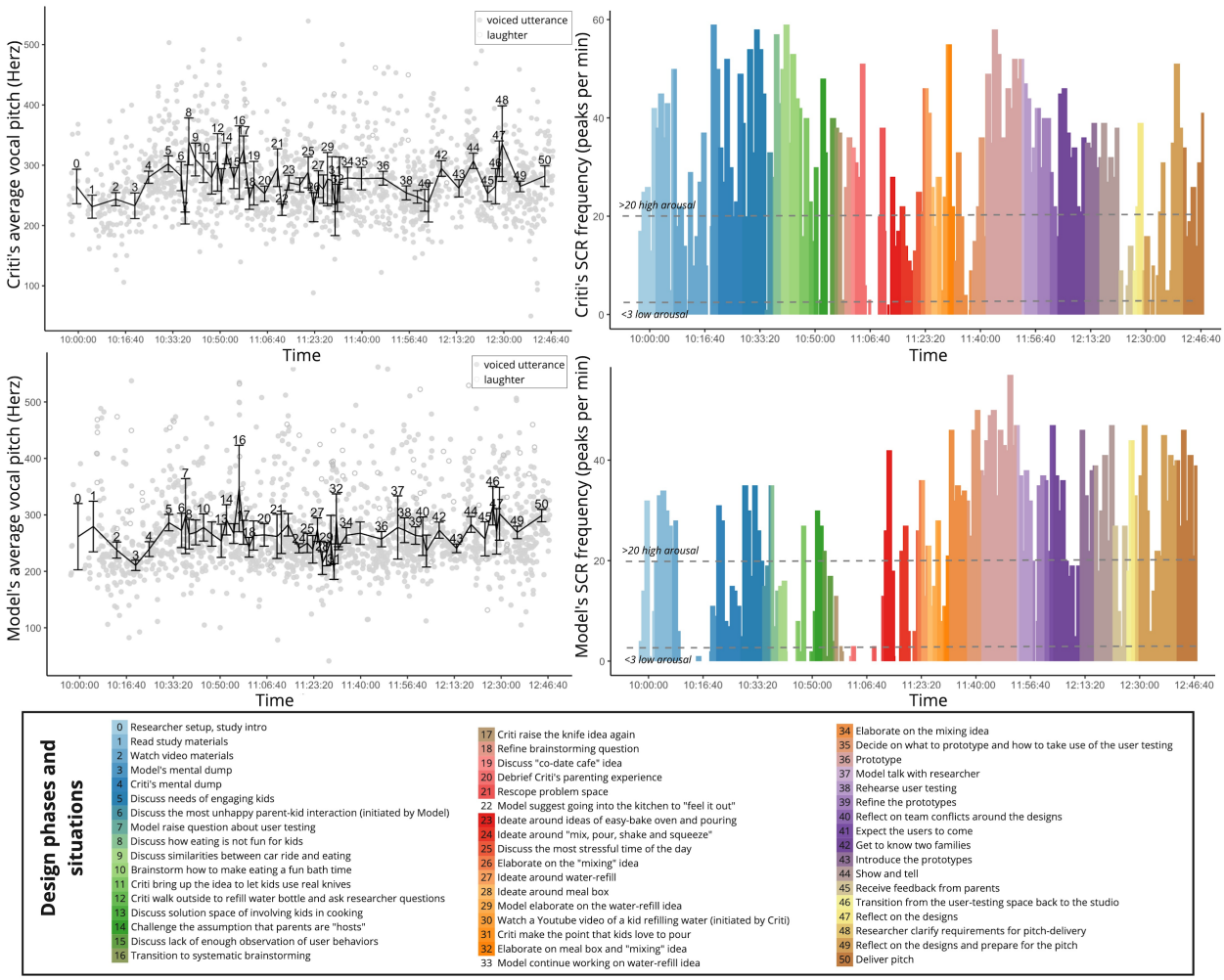


Figure 12. The three graphs here show the physiological data of all designers introduced in this study. The detailed description is to be added. As an example, in Analyte's case, the objective measures of vocal pitch (top-left) and SCR frequency (top-right) highly correlate (r=0.62, p=0.014) and also match well with Analyte's subjective report. For instance, Analyte reported his emotion did not change much except for the user-testing stage (segment 30 to 33), which is confirmed in both physio-measure profiles. In contrast, concordance isn't consistently found in other designers.