Evaluating the temporal dynamics of a structured experience: Real-time skin conductance and experience reconstruction measures

Leisure Sciences
An Interdisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ulsc20

Evaluating the Temporal Dynamics of a Structured Experience: Real-Time Skin Conductance and Experience Reconstruction Measures

Wim Strijbosch, Ondrej Mitas, Tessa van Blaricum, Olaf Vugts, Carolien Govers, Moniek Hover, John Gelissen & Marcel Bastiaansen

To cite this article: Wim Strijbosch, Ondrej Mitas, Tessa van Blaricum, Olaf Vugts, Carolien Govers, Moniek Hover, John Gelissen & Marcel Bastiaansen (2021): Evaluating the Temporal Dynamics of a Structured Experience: Real-Time Skin Conductance and Experience Reconstruction Measures, Leisure Sciences, DOI: 10.1080/01490400.2021.1967233

To link to this article: https://doi.org/10.1080/01490400.2021.1967233

Published online: 28 Aug 2021.

LEISURE SCIENCES
https://doi.org/10.1080/01490400.2021.1967233

Evaluating the Temporal Dynamics of a Structured Experience: Real-Time Skin Conductance and Experience Reconstruction Measures

Wim Strijbosch[a], Ondrej Mitas[b], Tessa van Blaricum[a], Olaf Vugts[c], Carolien Govers[c], Moniek Hover[a], John Gelissen[a,d], and Marcel Bastiaansen[a,d]

[a]Academy for Leisure & Events, Breda University of Applied Sciences, Breda, The Netherlands; [b]Academy for Tourism, Breda University of Applied Sciences, Breda, The Netherlands; [c]Efteling, Kaatsheuvel, The Netherlands; [d]Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands

ABSTRACT

Structured experience (SE) providers continuously evaluate and improve their experiential offerings to make them more memorable. Arguably, the temporal dynamics of the emotions in an experience have a crucial influence on its memorability. Traditional post-experience evaluation procedures tend to ignore these temporal dynamics, thus offering imprecise feedback for providers on exactly when and where to optimize their experiential offerings. In this paper, we use two methods as a tool for evaluating how closely the lived experience of a SE follows the experience as intended by the provider: real-time skin conductance (SC) and experience reconstruction measures (ERMs). We demonstrate that both SC and ERMs are significantly related to intended experience. This link was found to be stronger for later sections of the experience than for earlier sections. In addition, SC and ERMs appear to be useful tools to assess the effectiveness of design interventions, thus providing valuable feedback for SE providers.

ARTICLE HISTORY
Received 8 January 2021
Accepted 5 August 2021

KEYWORDS
Experience design; experience reconstruction; physiology; skin conductance; structured experience

Introduction

From the experiential perspective, leisure is often seen as a convergence of relative freedom and intrinsic motivation (Kleiber et al., 2011). Along these lines, leisure providers are seeking to improve and enhance experiences of leisure participants through carefully structuring their experience products (Scott et al., 2017). Duerden et al. (2015) conceptualize *structured experiences* (henceforth: SEs) as the objective encounters between participant and provider as well as the resulting subjective participant outcomes. More specifically, SEs are characterized as an ongoing process of co-creation between participant and provider, for which participants bring their relationships, thoughts, emotions, values, memories of past experiences and personal characteristics, and for which providers manipulate specific design elements with the intention to evoke an experience that is thought through from start to end (Duerden et al., 2015). Ellis and colleagues

CONTACT Wim Strijbosch strijbosch.w@buas.nl Academy for Leisure & Events, Breda University of Applied Sciences, Mgr. Hopmansstraat 2, Breda, 4817JS, The Netherlands.

(2019) have extended the conceptualization of SEs by building the theory of structured experience (TSE), which has been revised into TSE 2.0 following several empirical assessments (Ellis et al., 2020a, 2020b). In TSE, SEs are considered as "a planned invitation by an experience provider (i.e., a 'manipulated framework') for a heightened subjective state of motivation, attention and emotion to occur" (Ellis et al., 2019, p. 99). This invitation may then evoke two types of outcomes: immediate conscious experience and after-the-fact evaluations (Ellis et al., 2019, 2020b). Immediate conscious experience refers to several subjective states, such as immersion, absorption, engagement, or deep structured experience (DSE) (Ellis et al., 2019). Experience evaluations refer to after-the-fact evaluations subsequent to disengagement from the experience itself, which may consist of value judgments, affective responses, satisfaction/delight, and proclivity to recommend (Ellis et al., 2019, 2020b).

As Duerden et al. (2015) suggest, the SE framework is particularly useful for SE providers to first think about the outcomes they hope to achieve, and to subsequentially structure the individual SE elements with these outcomes in mind. At the same time, Duerden and colleagues (2015) acknowledge that the SE framework is not likely to account for *all* experiences of *all* leisure participants. Smit et al. (2020) note that experiences cannot be designed to create identical behavioral and neurological responses within each consumer. At best, a provider can structure the settings and surroundings with the *intention* to create similar effects on each consumer (Smit et al., 2020). However, to the best of our knowledge, the extent to which the actual experience of leisure participants can be in line with the intended experience of a provider has not been empirically evaluated in the literature to date.

Notwithstanding, the literature offers a wide variety of measurement scales to empirically evaluate experiences on a number of dimensions. For example, De Geus et al.'s (2016) Event Experience Scale measures the extent to which event visitors experience affective, cognitive, and physical engagement, as well as newness. Lee and Smith's (2015) Visitor Experience Scale measures the extent to which museum visitors experience entertainment, escapism, education, relationship development and culture identity-seeking. In the context of consumption experience, Klaus and Maklan's (2012) EXQ measures consumers' peace-of-mind and outcome focus, and asks consumers to evaluate moments-of-truth and various product features. In a similar vein, Parasuraman et al.'s (1988) SERVQUAL asks consumers to evaluate a service provider on its tangible products, reliability, responsiveness, assurance and empathy. Finally, for the context of SEs in particular, several scales have been developed to measure TSE-related constructs of engagement, absorption, immersion and DSE (see Ellis et al., 2020b for review). Using these scales, participants are typically surveyed after-the-fact and are then asked to aggregately evaluate their experience across the entire duration of their participation in a SE from start to end.

While the abovementioned scales could be a starting point for explicitly comparing intended experience with actual experience, a disadvantage of aggregative after-the-fact methods is that they disregard how experiences evolve over time. Recent accounts in the literature conceptualize experience as a temporally dynamic phenomenon of perceptual, cognitive, and affective processes (Bastiaansen et al., 2019; Stienmetz et al., 2021). The pattern and shape of these temporal dynamics have a strong influence on how experiences are evaluated and remembered (Fredrickson, 2000). Due to their aggregative nature, the abovementioned experience scales fail to capture the crucial information that lies in this temporal flux of emotions and perceptions (Bastiaansen et al., 2019; Stienmetz et al., 2021). As a consequence, such methods offer an incomplete understanding of experience and imprecise feedback for providers on exactly when SEs influence overall evaluations. In the present paper, we address these shortcomings by evaluating the SE of a musical theater show in a theme park in northwestern Europe through two types of measures that allow for capturing temporal dynamics: real-time measures of skin conductance and self-reported experience reconstruction measures.

Between lived and remembered experience

In the SE framework, SE refers to both the physical encounters between providers and participants, as well as the subjective states and outcomes on the side of the participant (Duerden et al., 2015). Jantzen (2013), however, argues that experience by definition is a psychological phenomenon, i.e., the subjective states under the SE framework. At best, the physical encounter between participant and provider provides the stage or context for these subjective states to take place and should not be considered as part of the psychological phenomenon itself. Separating the context from the psychological phenomenon is useful for subsequent discussions, as the term "experience" links together several other psychological concepts (i.e., lived and remembered experience; see below). For clarity, we henceforth thus refer to physical participant-provider encounters as *structured experience* (SE) and to the resulting psychological phenomenon simply as *experience*. Following Bastiaansen et al. (2019), the latter term has two components that are distinguished by their temporal occurrence: *lived experience* and *remembered experience*. *Lived experience* can be defined as an ongoing stream of consciousness, based on stimuli that are processed by our perceptual and affective systems (Jacobs, 2006). This continuous stream of experiencing is segmented into experiential episodes that have a clear beginning and ending through the activation of so-called mental models (Bastiaansen et al., 2019). An example is that the concept of 'playing soccer' segments your weekly soccer match off from whatever happened before or after.

Occasionally, an experiential episode will be stored into memory to form a *remembered experience* (Kahneman & Riis, 2005). Not everything we experience is automatically stored in memory (Zajchowski et al., 2017). Memorability, however, is an important asset for leisure providers, as experiences need to be remembered in the first place before they can develop into experiences with stronger impacts (i.e., meaningfulness or transformability) (Duerden et al., 2018) and before they lead to downstream behavior such as proclivity to recommend or revisit (Barnes et al., 2016). This begs the question of which elements in an experience make it more memorable.

A growing body of work suggests that the memorability of a leisure experience mostly depends on the emotions that are evoked in the lived experience (Bastiaansen et al., 2019; Del Bosque & San Martín, 2008; Duerden et al., 2018). This has been supported with sound empirical evidence in the field of general psychology that has studied the relationship between emotion and memory (see Holland & Kensinger, 2013 for review). In SEs in particular, emotion has been found to play an important role in determining engagement (Ellis et al., 2020a), which in turn has been argued to affect memorability as well (Scott et al., 2017). Therefore, we argue that emotionally engaging leisure consumers should be a core consideration in the design and development of SEs. As such, they can be of relevance in evaluation and optimizing SEs as well. For evaluation and optimization procedures of SEs, Bastiaansen et al. (2019) suggest that providers should first focus on determining which emotions should be felt when (the intended experience), and subsequently comparing the intended experience to the emotions that are actually experienced.

Emotions and how to measure them

In the emotion literature, emotions are generally understood as brief responses to stimuli that are seen as personally significant (Frijda, 1986). However, according to the most recent developments in the theory of emotion (Feldman Barrett, 2017), emotions are not a response to a stimulus, but a simulated anticipation of what that stimulus is going to be and how, given prior experience, the internal milieu (i.e., mental and bodily states) can best respond to this stimulus in terms of body budget. Through interoception, these mental and bodily states are available to consciousness through the dimensions of emotional valence and arousal, together referred to as core affect (Russell, 2003). Previously stored mental models then allow for an interpretation of these dimensions in terms of discrete emotion labels used in daily language, such as anger, fear and joy. While mental models of emotions are culturally determined, dimensions of emotional valence and arousal are a human universal (Feldman Barrett, 2017).

Emotions are expressed phenomenologically, behaviorally and physiologically and can hence be measured along these dimensions as well (Mauss & Robinson, 2009). In the leisure literature, emotions are most commonly measured phenomenologically through the use of self-report (Li et al., 2015), either by tapping into the dimensions of valence and arousal or by asking participants to report on discrete emotions. This approach has been criticized, as it generally measures participant's cognitive evaluation of an emotion, rather than the emotion itself. Also, self-report is generally conducted after-the-fact, thus measuring remembered emotion rather than lived emotion (Li et al., 2015). As noted by Kahneman and Riis (2005), lived experience uses different sources of input than remembered experience. Lived experience is based on (un)conscious sources of input in the here and now, whereas remembered experience is based on after-the-fact interpretations through several heuristics, such as recency effects and peak-and-end-mechanisms (see e.g., Strijbosch et al., 2021).

Several recent accounts in the literature suggest electrophysiological measures as a promising tool for studying the emotions over the course of an experience (Bastiaansen et al., 2019; Li et al., 2015). According to Birenboim et al. (2019), electrophysiological equipment can advantageously record real-time physiological signals (such as heart rate, body temperature and skin conductance) that form a more objective measure of emotions than phenomenological or behavioral assessments. In addition, electrophysiological equipment allows for measuring these signals with sub-second precision, resulting in a detailed temporal profile of the emotional responses during an experience. Electrophysiology is able to capture emotions unobtrusively, thereby not affecting the experience itself. Electrophysiological measures have certain drawbacks, too (Birenboim et al., 2019). First, the rather sensitive equipment is prone to measurement artifacts (i.e., peaks in the recorded signal that result from sources other than from bodily responses). Also, electrophysiology asks for maximal experimental controllability, so as to control for environmental factors that are crucial for data interpretation. Furthermore, while experience is essentially a product of human consciousness (Jacobs, 2006), electrophysiological measures of emotion capture bodily processes, which are of a different nature and not necessarily convergent with measures that tap into consciousness (Mauss & Robinson, 2009). To obtain a more comprehensive measurement of emotions, it is therefore advisable to use electrophysiology together with additional phenomenological or behavioral methods (Hosany et al., 2020). Suggested phenomenological methods that are able to capture temporal dynamics as well are experience sampling methods or experience reconstruction methods. Experience sampling methods use immediate, real-time, self-reported measures of lived experience as it is taking place (see e.g., Hektner et al. 2007). Experience reconstruction methods retrospectively reconstruct lived experience from memory by recording self-reported measures on segments of the lived experience directly after it has ended (see e.g., Strijbosch et al., 2019, 2021).

An electrophysiological measure that has gained particular interest in the leisure literature is that of skin conductance (SC), which after applying the proper mathematical procedures forms a temporally well-aligned proxy for emotional arousal to be measured in SEs (Bastiaansen et al., 2020; Birenboim et al., 2019; Kim & Fesenmaier, 2015; Li, 2021; Mitas et al., 2020; Strijbosch et al., 2021; Van Bendegom et al., in press). SC does not discriminate between positively or negatively valenced emotions (Boucsein, 2012). However, much like other electrophysiological measures, SC has the advantage of precisely indicating when participants are experiencing an emotionally arousing response. In maximally controlled settings, these responses can be linked to environmental cues, such as the touchpoints in SEs. In the studies mentioned above, SC has been used to give an idea of the emotional arousal during SEs that are both highly structured (such as roller coasters, theme parks, guided museum tours and theater shows (see e.g. Bastiaansen et al., 2020; Li, 2021; Mitas et al., 2020; Strijbosch et al., 2021; Van Bengedom et al., in press)) and less structured (such as city walks (see e.g. Birenboim et al., 2019; Kim & Fesenmaier, 2015). In these contexts, SC has been studied in relation to such evaluation measures as proclivity to recommend (Mitas et al., 2020; Strijbosch et al., 2021), appreciation (Strijbosch et al., 2021) and assessments of overall emotional valence and arousal (Van Bendegom et al., in press; Strijbosch et al., 2021). In the present work we use SC to compare the time course of emotional arousal over the lived experience of leisure participants with the time course of emotional arousal as intended by the designers of a SE. We thus address the question whether SC can be used to evaluate how closely the lived experience follows the intended experience.

The present study

In the present paper, we explore SC as a tool for evaluating the SE of a musical theater show in a theme park in northwestern Europe. To address the phenomenological nature of experience as a product of human consciousness, we complement this method with an experience reconstruction method adapted from Strijbosch et al. (2019). This method surveys participants directly *after* the experience on their emotional valence and arousal during (either predefined or participant-defined) episodes within the lived experience. Ratings of valence and arousal during each segment are then used to create temporal profiles of emotional valence and arousal during the lived experience, thus complementing the SC profiles captured during the lived experience real-time.

The 75-minute musical theater show was selected to ensure high environmental controllability and to address the susceptibility of the electrophysiological equipment to motion artifacts that otherwise result from excessive movement of participants. Also, due to its tightly scripted nature, the show provides very precise temporal information about which stimuli are experienced at which moment in time. This makes the show an excellent setting for comparing the lived experience of participants with the intended experience. The show meets the definition of a leisure experience, as it is freely chosen and offers participants no other explicit rewards besides attending the show itself.

To assess the usefulness of SC and experience reconstruction measures (ERMs) (for clarity, we henceforth refer to SC and ERMs together as *lived experience measures*) as tools for evaluating SEs, we compare them against the intended experience as established by the provider. As SEs are designed to deliver experiences in an intentional manner, we hypothesize that:

(H1) The time course of lived experience measures during the musical theater show is related to the time course of the intended experience for that same show.

In addition, to see whether our lived experience measures can be used as a tool for SE *optimization*, we established whether they would covary with modifications to the design of the theater show. Therefore, the show was offered in two versions: an *original* and an *adapted version* in which 2 of the 17 original show scenes (scene 1 and 7) were redesigned. Scene 1 was redesigned by letting a show interact with audience while walking through the seating area instead of from the central stage. Scene 7 was redesigned by changing a sequence of hypnosis, dance and pyrotechnic effects into an intense scene where all these things happened at once, instead of in sequence. Both interventions were based on creative decisions by the show's creative development team to make the scenes emotionally more arousing, so as to better align the respective scenes with the intended experience. Particularly, the role of these interventions in the current study was to see whether our lived experience measures could establish the effectiveness of the interventions. Therefore, we hypothesize that:

(H2) The time course of lived experience measures during the adapted version of the musical theater show more closely follows the time course of the intended experience, than the time course of lived experience measures during the show's original version.

(H3) Scene 1 of the adapted show induces more emotional arousal than in the original show.

(H4) Scene 7 of the adapted show induces more emotional arousal than in the original show.

Materials and methods

Participants

Participants were selected from the research panel of the theme park resort that hosts the musical theater show. Panel members received an email with information about the study, for which they could sign up voluntarily. From all subscriptions, 89 participants were selected to compose a quota sample matched with the target group of the show, based on age and family composition. The target group was defined as families with two adults and two children above the age of 6. These families were then invited to visit the show on a set date in either January–February or May–June 2019, for which the theme park resort offered them free tickets as an incentive to stimulate participation. Of the participating families, only the adults were selected to take part in the experiment. They were informed about the set-up of the study and subsequently gave their written informed consent in accordance with the Declaration of Helsinki.

The 89 participants were divided over two groups: one group that watched the original version of the show in January or February ($n = 39$), and one group that watched an adapted version of the show in May or June ($n = 50$). Participants were unaware of differences between the two versions and subscribed based on their availability in the mentioned periods. Of the 39 participants for the original show, the physiological data of 6 participants turned out to be either missing due to equipment failure or to contain excessive movement artifacts in the physiological signal (see the *Data analysis* section for details). These participants were excluded from further analysis, resulting in a final group of 33 participants ($M_{age} = 43.09$; $SD_{age} = 7.26$; 17 males/16 females). Similarly, of the 50 participants for the adapted show, 7 participants were excluded, resulting in a final group of 43 participants ($M_{age} = 45.21$; $SD_{age} = 8.04$; 17 males/26 females). This group consisted of more females compared to the first group, but this difference was not significant ($\chi^2{}_1 = 1.084$; $p = 0.356$), neither was the difference in age ($t_{73} = -1.174$; $p = 0.244$). All participants were Dutch nationals. Note that although small sample sizes are common in psychophysiological studies, our final sample of 76 participants is substantially larger than the typical $n = 20$–25 in traditional psychophysiological laboratory studies (Bastiaansen et al., 2020).

Structured experience

The SE consisted of a 75-minute musical theater show that was performed weekly in the theme park resort. The show's concept was drawn from the theme of a carousel of life. Relying mostly on mimicry, music and visual spectacle, the show was almost entirely performed without spoken language. Before the beginning of the show, two show hosts select an (informed) guest from the audience who becomes the protagonist of the show (scene 1). The stage is then magically changed into a carousel of time. The hosts take the protagonist back to his childhood, where he meets his prospective girlfriend and wife (scene 2, 3, 4 and 5). As the show hosts make time go faster, the protagonist wants to travel the world and sets out on a voyage to explore foreign and exotic cultures (scene 6, 7 and 8). After receiving several letters from his female friend back home, the protagonist returns home, where the two meet up again (scene 9). They get married (scene 10) and become parents of a baby girl (scene 11). As the young parents grow older, they end up in the hassles of parenting and working life (scene 12). At one point, the woman cannot keep up with her husband and child anymore, suddenly faints, and eventually dies (scene 13), leaving her husband and child behind. Eventually, the child's resilience drags her father out of depression as their sorrows are washed away by a magical water curtain (scene 14). Both pick up their life and the protagonist grows into being a grandfather as his daughter gives birth to a baby boy (scene 15, 16). The baby boy grows up to the same age as his grandfather in scene 2 and meets a female friend of his own, starting his own ride on the carousel of life. The show ends with all actors on stage, accompanied by bombastic music, fireworks and a round of applause (scene 17).

The two versions of the show differed in terms of scene 1 and scene 7. Originally, scene 1 consisted of one of the show hosts calling various audience members to the stage to select the protagonist through a series of playful assignments. In the adapted version, the host walked through the audience and conducted those assignments in the seating area, creating a more arousing experience by limiting the distance between the host and the audience members. Scene 7 originally consisted of a fakir that hypnotized the protagonist for a dance with a snake, followed by several pyrotechnic effects. For the adapted version, this scene was modified so that the hypnosis, dance and pyrotechnic effects happened all at once to increase the intensity of the scene. For the remainder of the show, the two versions were identical.

Establishing the intended experience

Before collecting data from the participants, four members of the theater show's creative development team (two producers, the director and the composer) were individually interviewed to determine how they intended visitors' emotional arousal to evolve over the time course of the show, using a semi-structured interview technique. First, interviewees were asked to talk about their role in developing the show. Then, they were asked to talk through the timeline of the whole show and to reflect on what emotions they intended for guests to feel during each of the 17 scenes. Near the end of the interview, the interviewees were provided with a diagram with the timeline of the show on the horizontal axis and emotional arousal on the vertical axis, using a 9-point scale for emotional arousal (Bradley & Lang, 1994) (see Figure 1). They were then asked to draw a temporal profile of the intended emotional arousal for the whole show (for clarity, we henceforth refer to this temporal profile as the *intended experience*). We chose to determine the intended experience in terms of emotional arousal, as emotional arousal is a prominent element in the dramatic arc, a tool often used in story-driven experiences such as the current theater show (Calvi & Hover, 2016). The dramatic arc is a storytelling structure that follows an increase in emotional tension or suspense as the story unfolds, peaking at the climax and then returning to normal as the story ends (Calvi & Hover, 2016). Also, arousal is the dimension of emotion that is often linked to SC (Boucsein, 2012).

The four "analogue" time series of emotional arousal provided by the members of the creative development team were digitized and linked to the onset/offset of the events and scenes (see Figure 1). When an "analogue" profile exceeded the 9-point scale as originally provided, the whole profile was first digitized and then recalculated into a 9-point scale. The four individual profiles were averaged to establish the intended experience for the creative team as a whole, which was then used for further analyses.

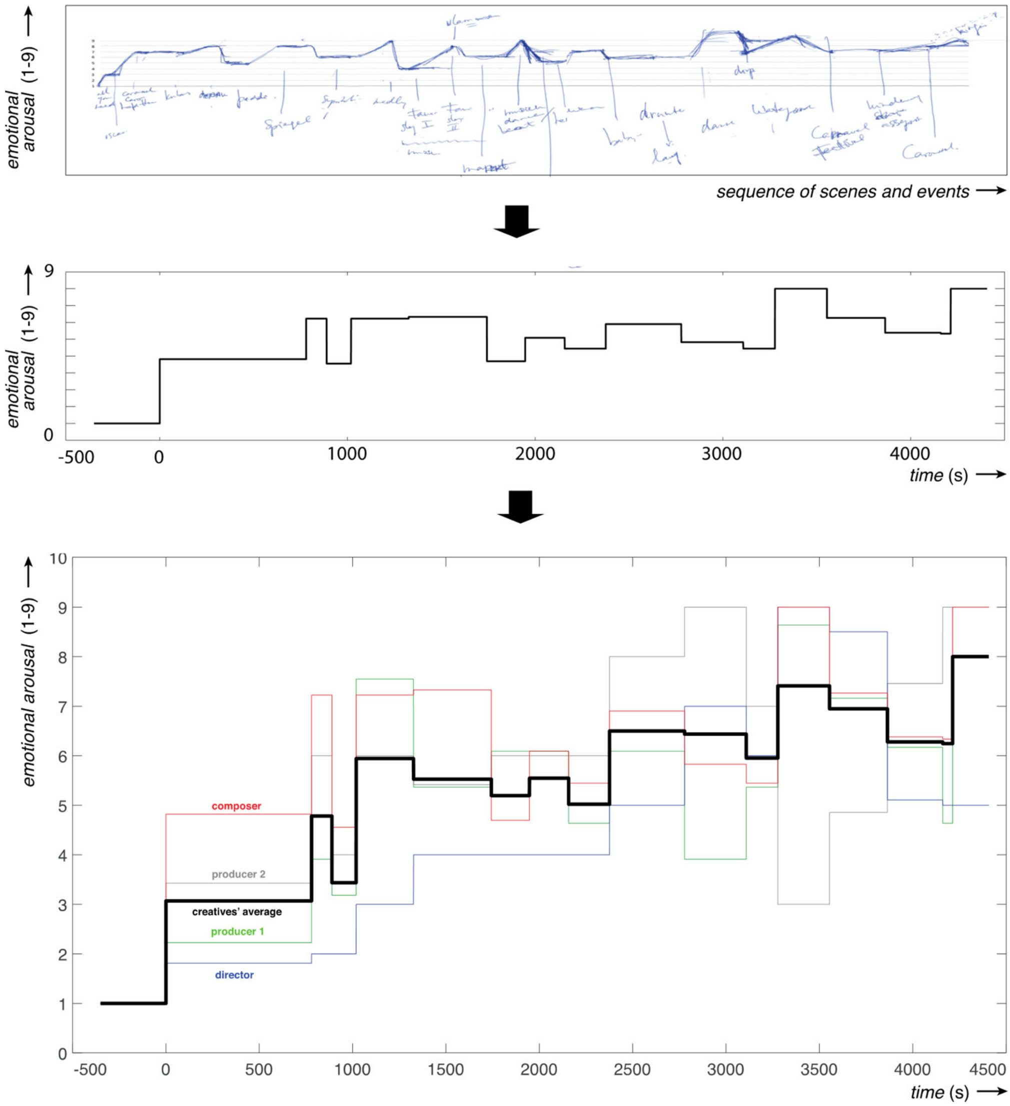


Figure 1. The top and middle sections of the figure illustrate how the hand-drawn intended experience from the interview with one of the creative developers was transformed into a digital version. The bottom section of the figure shows the different intended experiences from the individual creative developers and the average intended experience over that of the four creatives.

Design and procedure

Data were collected over 11 performances of the show (6 original in January–February 2019; 5 adapted in May–June 2019). After having obtained written informed consent, the experimenter put an Empatica E4 wristband on the wrist of the participants' non-dominant hand. Participants were instructed to sit in the theater as relaxed as possible, and to not touch or move the wristband during the show. They were then seated in the theater and watched the show together with other nonparticipating audience members so as to create an ecologically valid setting. To guarantee independence of each individual's experience, couples were seated so as to be physically separated. After the show, wristbands were removed from the participants, who were then given a post-experience questionnaire.

Data collection

Physiological data

SC data were recorded with Empatica E4 wristbands (Empatica Inc., USA). While various alternatives exist (e.g., the Shimmer3 GSR+ (Shimmer Research Ltd, Ireland) and the EdaMove 4 (movisens GmbH, Germany)), Empatica E4s were chosen for their user-friendliness (both for participant and experimenter), the inclusion of a button to mark a temporal event (e.g., the beginning of the show) as well as the fact that the user has access to the raw, untreated data. Physiological data were continuously sampled at a frequency of 4 Hz and were stored on the Empatica for further off-line processing. Measurement started when participants began to wear the Empaticas and continued until the experimenter removed the wristbands. Time was also recorded on the Empaticas. In order to obtain an indicative time alignment between the recordings of SC data and onsets/offsets of each of the scenes, the experimenter wrote down time stamps of the onset/offset of each of the show's performance, as well as around the onset of the 17 different scenes. For this, the experimenter used a mobile telephone with the time-zone synchronized Alarm Clock Pro app (iHandy Ltd., Hong Kong). As soon as each event related to the scene onsets occurred, the experimenter noted the exact time that was displayed on the mobile phone at that moment.

Post-experience questionnaire

In the post-experience questionnaire, participants were asked to indicate their emotional valence and arousal for each of the 17 show scenes, using 5-point scales adapted from Bradley and Lang (1994). Scenes were referred to through the use of a picture that was considered visually representative for the whole scene, followed by the questions "To what extent did you feel positive or negative during the scene as depicted by the photo?" and "To what extent did you feel calm or excited during the scene as depicted by the photo?" using the same 5-point scales as mentioned above. We explicitly chose to use pictures rather than verbal labels, so as to not influence participants with interpretations that might affect their answers (i.e., labeling scene 12 as "the stress scene" might prime participants to evaluate the scene as negatively valenced because of associations with the word "stress").

Data analysis

Pre-processing of skin conductance data

SC data were extracted from the Empaticas, stored on a PC and imported into MATLAB (MathWorks, USA) for further analysis. First, SC data were precisely time-synchronized with the onset/offset of the show, using the time denotations from the experimenter. SC data corresponding to the length of the performances were then extracted from the recordings and analyzed according to the procedure proposed in

Bastiaansen et al. (2020) as briefly described below. Data from 13 participants were excluded from further analysis because of equipment failure, or because their physiological data contained too many artifacts. In the remaining 76 participants, we removed any evident motion artifact detected by Bastiaansen et al.'s (2020) artifact detection and removal algorithm. SC data were then subjected to a continuous deconvolution analysis to separate the signal into its tonic (slow and long-term changes) and phasic (rapid short-term changes) drivers. The phasic driver most accurately corresponds with emotion-related central nervous system responses and thus serves as a proxy for emotional arousal (Boucsein, 2012). Continuous deconvolution was performed with the open-source MATLAB toolbox of Ledalab (Benedek & Kaernbach, 2010). The resulting phasic driver was then standardized using a z-transformation for comparability across participants.

To translate the continuous phasic driver into per-scene values, per participant, for each scene we took the maximum phasic value for that scene (henceforth: peak SCR) as well as the average of the phasic values for that scene (henceforth: average SCR), using the onset/offset times of the 17 different scenes. For most scenes, peak and average SCR were calculated from the phasic driver of the entire scene. Some of the scenes, however, included intense physical audience participation (scene 6) or consistently evoked long-lasting episodes of applause (scene 8, 11, 15 and 17). These movements generated excessive motion artifacts for all participants across all performances of the show. For these particular scenes, time segments containing such motion artifacts were omitted when calculating peak and average SCR to avoid contamination of these measures (for a visual representation of this procedure, see Figure 2). Peak and average SCR per scene served as the input for further analysis.

Pre-processing of reconstructed experience measures

The post-experience questionnaire resulted in 17 per-scene emotional valence and arousal ratings: one for each scene. Using the onset/offset times of the individual scenes, a temporal profile of emotional valence and arousal ratings was reconstructed (see Figure 3). These profiles were used as a representation of the lived experience as measured through self-report.

Statistical analyses

To test hypotheses H1 and H2, we used Singer and Willett's (2003) multilevel model for change. This approach uses a multilevel model framework to see how repeated measures (level 1) nested within participants (level 2) change over time. A level-1 predictor modeling time is included as a main effect, with random intercepts for examining differences in initial status in the outcome variable and a random slope for evaluating whether there are differences in change trajectories over time between participants. Hypothesized covariates are then entered into the model to see whether differences in initial status and change trajectories between participants can be explained by level-2 predictors such as age, gender or an experimental manipulation.

In our approach, we used Stata (StataCorp LLC, USA) to estimate the multilevel models following the strategy as described above, using Full Maximum Likelihood to fit

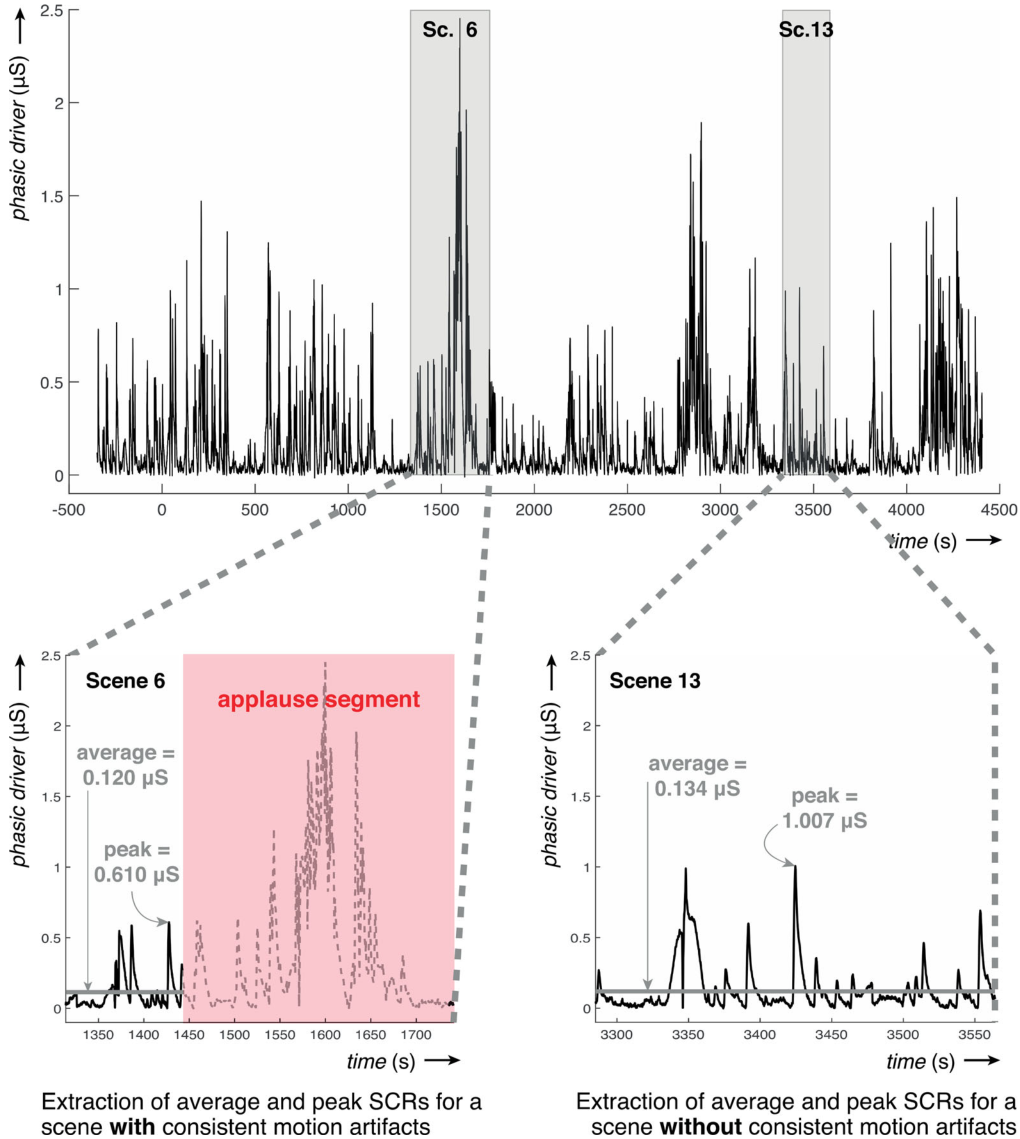


Figure 2. Extraction of average and peak SCRs for scenes with and without consistent motion artifacts. The top figure displays the phasic driver of the full show for one participant. Bottom figures are sections from this driver corresponding with the onsets/offsets of scene 6 (left) and scene 13 (right). Scene 13 has no applause segments, and average and peak SCRs are therefore based on the scene's full signal. Scene 6, however, ends with an extensive segment of applause (marked with the red box). For this scene, the applause segment was left out, and average and peak SCR were extracted from the remaining signal.

the models. In correspondence with the 17 different show scenes, in this study, repeated measures consisted of 17 values within a participant for each of the four lived experience measures (self-reported valence, self-reported arousal, average SCR, and peak SCR). In our models, the level-1 predictor of time thus consisted of the scene numbers ranging from 1 to 17. To see whether random slopes of this predictor would improve the model fit, we tested the model with random slopes for the time predictor against a model in which slopes for the time predictor were fixed to zero.

Our model-building strategy is as follows. H1 states that intended experience predicts the lived experience measures. For testing H1, in a first model (Model 1), we included two level-1 predictors to predict the four lived experience measures: *scene number* (our time predictor) and *intended experience*; the latter as a time-varying covariate that is nested within participants. Much like the time predictor, *intended experience* consisted of 17 values (1 value per scene), as derived from the digitized temporal profile in Figure 1. We included the cross-product between the intended experience and scene number, and allowed the effect of the intended experience predictor to vary over time. This yields the following main Model 1 equation for all lived experience measures:

$$\begin{aligned} LIVED.\mathrm{EXP}_{ij} = {} & [\gamma_{00} + \gamma_{10} SCENE.NUMBER + \gamma_{20} INTENDED.\mathrm{EXP} \\ & + \gamma_{30} SCENE.NUMBER \times INTENDED.\mathrm{EXP}] \\ & + [\zeta_{0i} + \zeta_{1i} SCENE.NUMBER + \varepsilon_{ij}] \end{aligned}$$

with γ_{00} being the initial lived experience when intended experience equals 0, γ_{10} being the effect of scene number on lived experience when intended experience equals 0, γ_{20} being the effect of intended experience on lived experience when scene number equals 0 and γ_{30} being the effect of intended experience on lived experience as the scenes progress during the show. ζ_{0i} and ζ_{1i} are between-person residuals in lived experience at the beginning of the show and of the effect of scene number, respectively, which are assumed to be bivariate normal with mean 0, unknown variances σ_0^2 and σ_1^2, and unkown covariance σ_{01}^2. ε_{ij} is the within-person residual variance which is assumed to be normally distributed with mean 0 and unknown variance σ_ε^2.

To test H2, we expanded our main Model 1 by adding the level-2 covariate of *show version* (original versus adapted), both as a direct effect only while controlling for other variables in the model (Model 2) or in interaction with intended experience and scene number (i.e., a model with two-way and three-way interactions), yielding Model 3. Model 2 allowed us to investigate whether differences between participants were directly due to the show's interventions and Model 3 to see whether the time-varying effect of intended experience on lived experience was different between show versions. For hypotheses H3 and H4, we used the same outcome variables as those under the analysis strategy for H1 and H2. As the hypotheses predict that the lived experience is different following interventions in two scenes, in this model, we included scene numbers over time as our main level-1 predictor and show version as our level-2 (between-persons) predictor. To see whether the lived experience for scene 1 and scene 7 was different for the adapted as compared to the original version of the show, we employed two planned comparisons.

For all models, we report the unstandardized regression coefficients, standard errors, and *p*-values of the fixed effects, as well as the estimates and standard errors of the random effects. Likelihood Ratio Tests (LRTs) and the Bayesian Information Criterion (BIC) are reported for evaluating whether model fit improved when covariates were added for explaining differences in lived experience outcomes. Concerning effect size measures, we follow Lorah (2018) and report Cohen's f^2 (Cohen, 1992), which is based on Snijders and Bosker's R^2 related to variance explained at the within-person level for the overall model (Snijders & Bosker, 2012). For the planned comparisons, we report the contrast-specific χ^2-values and accompanying p-values.

Results

Grand average temporal profiles of self-reported valence and arousal across all participants are reported in Figure 3. On average, self-reported valence across all per-scene ratings and all participants was mildly positive for both the original ($M = 3.857$, $SD = 0.771$) and the adapted version of the show ($M = 3.759$, $SD = 0.790$). Likewise, participants reported the show as being mildly arousing, on average, for both the original ($M = 3.346$, $SD = 0.925$) and the adapted version ($M = 3.400$, $SD = 0.814$). For both versions of the show, self-reported valence and arousal ranged from the lowest (1) to the highest (5) ratings. This finding indicates that apart from the grand averages, some of the scenes could trigger the extremes of emotional valence and arousal.

Although per-scene SCR values offer some idea of the distribution of these physiological measures, they cannot be compared against an absolute scale as with self-reported valence and arousal. For the physiological data, the temporal patterns and between-scene differences are therefore more meaningful than the individual SCR values. For an interpretation of these patterns, we refer to Figure 4, which displays the grand averages across participants of average and peak SCR per scene.

Predicting lived experience from the intended experience

As seen in Figure 3, the intended experience followed a steep, ramp-like pattern, whereas average self-reported arousal increased more slowly over time and self-reported valence merely fluctuated around a level of mildly positive valence. Figure 4 shows that unlike the intended experience, average and peak SCR followed a more wave-like pattern that begins relatively high, goes up and down in 2 or 3 waves and eventually ends with a high value.

H1 states that the time course of the intended experience is related to the time course of the lived experience measures. The evidence in Table 1 shows that all four lived experience measures could be significantly predicted from the intended experience at time = 0, as in Model 1 intended experience was found to be a significant main effect for self-reported valence and arousal, and average and peak SCR. Also, for all four lived experience measures, intended experience formed a significant positive interaction effect with scene number (our predictor of time), indicating that intended experience was a time-varying covariate. This means that the prediction effect of intended experience on the lived experience measures becomes stronger as the show progresses over time. In other words, the effect was stronger for later as compared to earlier scenes of the show. Guidelines for interpretation of Cohen's f^2 (Cohen, 1992) indicate that for the overall model with scene number, intended experience and the interaction between these variables, effect sizes are small-medium for self-reported valence, medium for self-reported arousal, and small for both average and peak SCR. Together, these results provided support for H1.

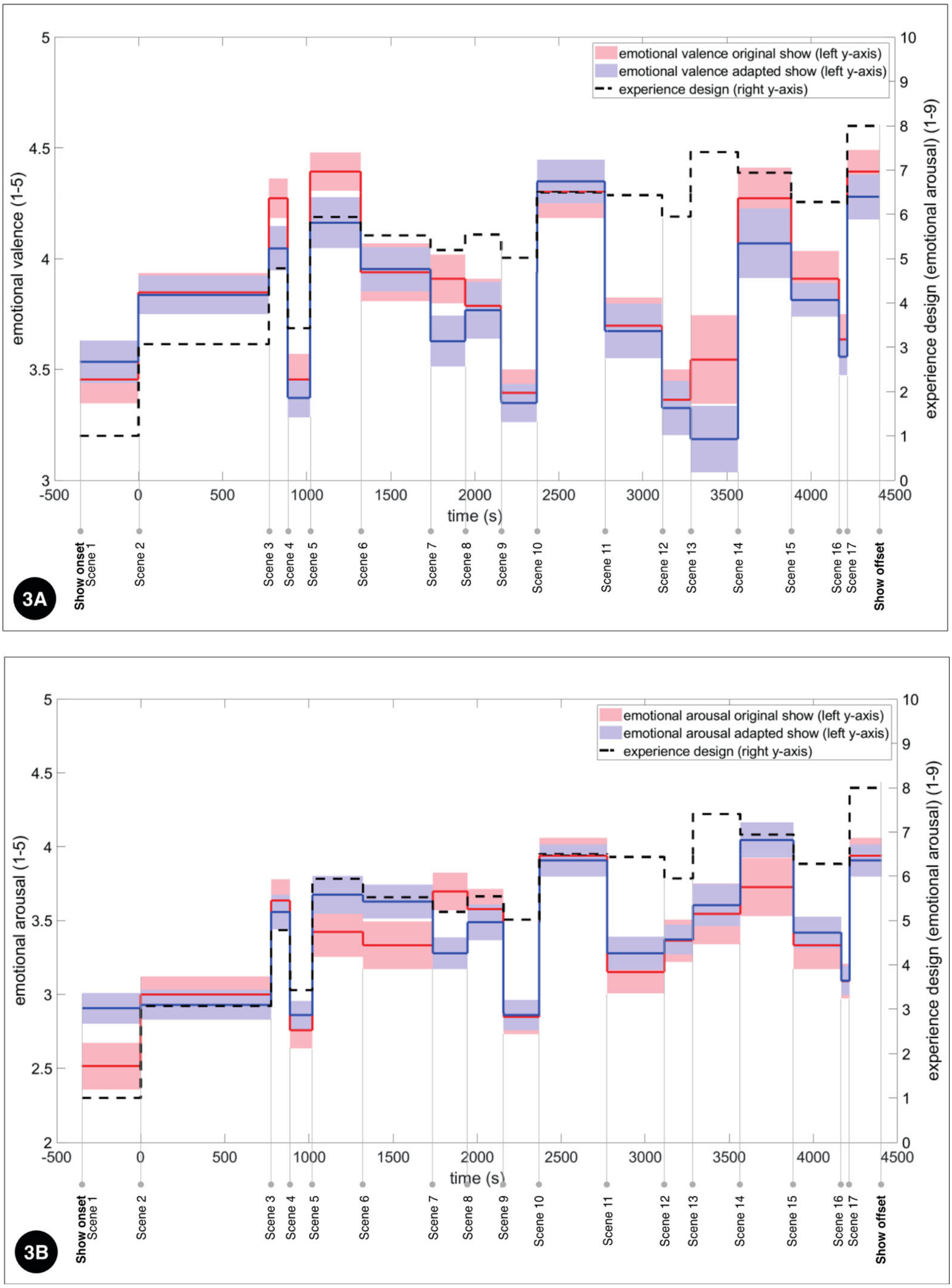
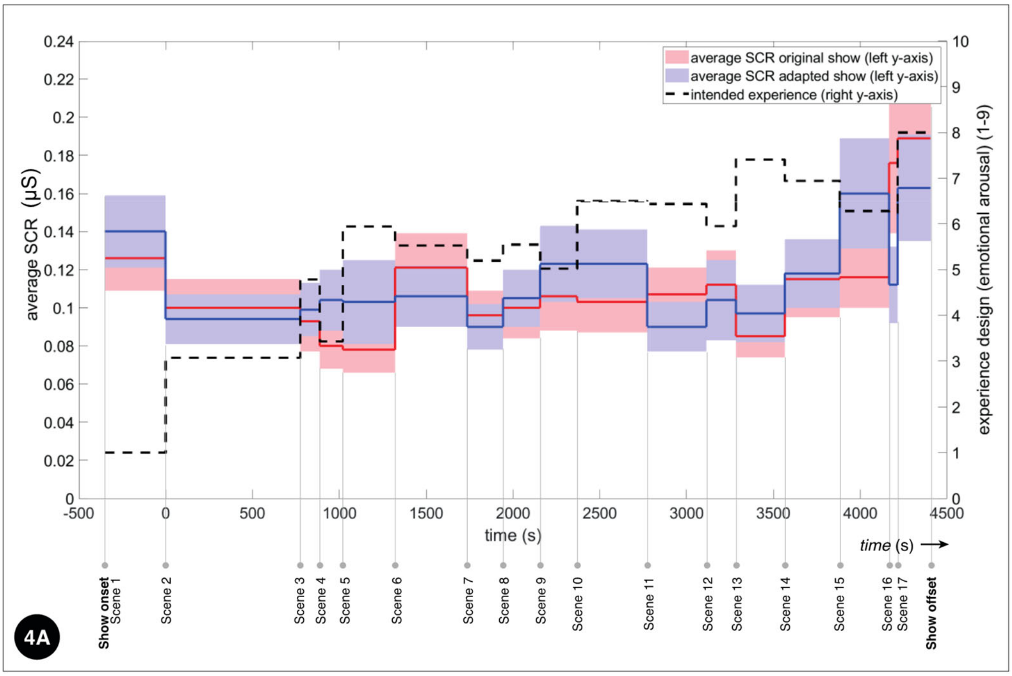


Figure 3. The grand average of the self-reported emotional valence (Figure 3A, top) and arousal profiles (Figure 3B, bottom) over all participants for both the show's original (indicated in red) and adapted version (indicated in blue). Shaded areas reflect the standard error of the mean. Emotional valence and arousal profiles follow the y-axis on the left. Intended experience (indicated with a dashed line) follows the y-axis on the right.



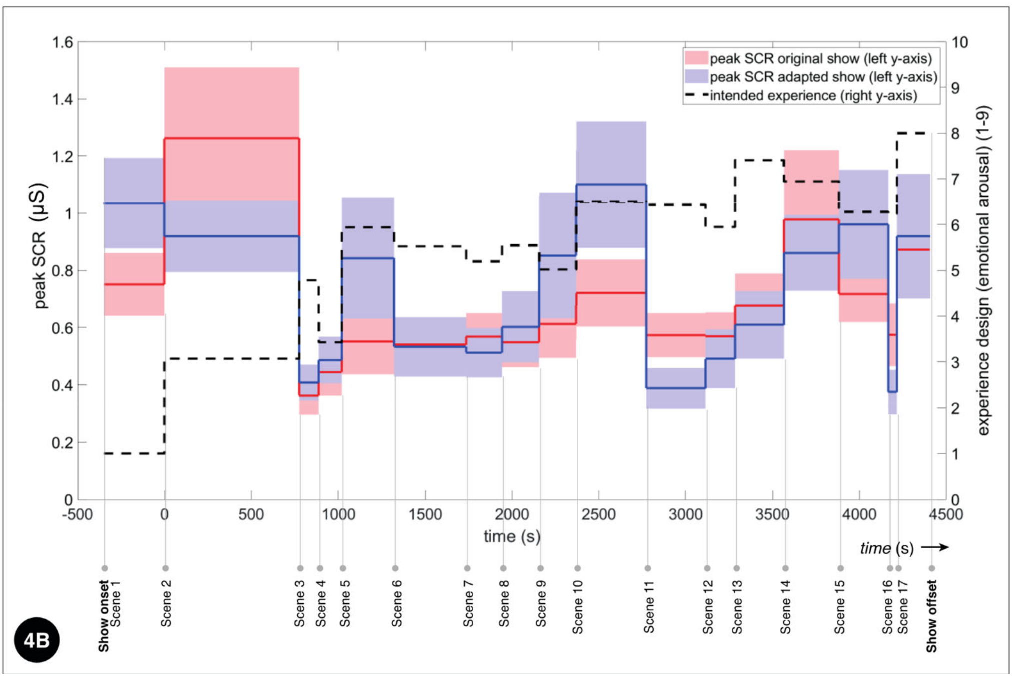


Figure 4. The grand average of average SCR (Figure 4A, top) and peak SCR per scene (Figure 4B, bottom) over all participants for both the show's original (indicated in red) and the adapted version (indicated in blue). Shaded areas reflect the standard error of the mean. Average and peak SCR profiles follow the y-axis on the left. Intended experience (indicated with a dashed line) follows the y-axis on the right.

Table 1. Model fits and parameters for the multilevel models for change under hypotheses H1.

	Dependent variables			
Independent variables and model indices	Self-reported valence	Self-reported arousal	Average SCR	Peak SCR
Model 1				
Intraclass correlation coefficient	0.141 (*SE* = 0.028)	0.218 (*SE* = 0.040)	0.601 (*SE* = 0.046)	0.346 (*SE* = 0.047)
Within-person fixed effect estimates				
Scene number	−0.166* (*SE* = 0.021)	−0.137* (*SE* = 0.020)	−0.003 (*SE* = 0.002)	−0.080* (*SE* = 0.020)
Intended experience	0.157* (*SE* = 0.023)	0.244* (*SE* = 0.022)	−0.014* (*SE* = 0.002)	−0.086* (*SE* = 0.022)
Scene number × Intended experience	0.017* (*SE* = 0.003)	0.013* (*SE* = 0.003)	0.001* (*SE* < 0.001)	0.014* (*SE* = 0.003)
(intercept)	3.504* (*SE* = 0.098)	2.514* (*SE* = 0.100)	0.146* (*SE* = 0.012)	1.095* (*SE* = 0.107)
Random effect estimates				
Variance of scene number	n/a (see *Note*)	<0.001 (*SE* < 0.001)	<0.001 (*SE* < 0.001)	<0.001 (*SE* < 0.001)
Variance of intercept	0.080 (*SE* = 0.018)	0.132 (*SE* = 0.031)	0.006 (*SE* = 0.001)	0.254 (*SE* = 0.051)
Residual variance	0.486 (*SE* = 0.020)	0.472 (*SE* = 0.020)	0.004 (*SE* < 0.001)	0.459 (*SE* = 0.019)
Goodness-of-fit indices				
BIC	2878.846	2914.840	−3005.196	2928.538
LRT against null model (χ^2_4)	117.25*	253.69*	214.52*	41.86*
Effect sizes				
Cohen's f^2	0.082	0.158	0.023	0.015
Snijders & Bosker's R^2 (within-person level)	0.076	0.137	0.023	0.015

Note. SE refers to standard error of the mean. LRTs of the models with random slopes for *scene number* against models with fixed slopes for *scene number* were significant for self-reported arousal ($\chi^2_1 = 8.43$; $p = 0.004$), average SCR ($\chi^2_1 = 135.32$; $p < 0.001$) and peak SCR ($\chi^2_1 = 10.94$; $p < 0.001$), but not for self-reported valence ($\chi^2_1 = 0.13$; $p = 0.716$). Therefore, we allowed random slopes for models with self-reported arousal, average SCR and peak SCR, but not for the model with self-reported valence.
*$p < 0.001$.

Table 2. Planned comparisons under hypotheses H3 and H4.

Independent variables and model indices	Dependent variables			
	Self-reported valence	Self-reported arousal	Average SCR	Peak SCR
Intraclass correlation coefficient	0.167 ($SE = 0.030$)	0.269 ($SE = 0.038$)	0.647 ($SE = 0.039$)	0.404 ($SE = 0.044$)
Goodness-of-fit indices				
BIC	2883.086	3010.632	−2744.299	3035.454
LRT against null model (χ^2_{33})	327.93**	365.66**	161.38**	142.70**
Planned comparisons				
Scene 1 adapted vs. original (χ^2_1)	0.25	4.78	0.26	2.10
Scene 7 adapted vs. original (χ^2_1)	3.00	5.43*	0.04	0.08

Note. For the planned comparisons, *p*-values have been adjusted following a Bonferroni correction. SE refers to standard error of the mean.
*$p < 0.05$.
**$p < 0.001$.

To test H2, stating that the lived experience measures of the adapted show follow the intended experience more closely than those of the original show, we expanded Model 1 by adding show version as a predictor to the model, both as a direct controlled effect (Model 2) and in interaction with other predictors (Model 3). Adding show version to the main model, however, did not lead to a better model fit for any of the lived experience measures (for Model 2: self-reported valence: $\chi^2_1 = 1.66$, $p = 0.198$; self-reported arousal: $\chi^2_1 = 0.51$, $p = 0.524$; average SCR: $\chi^2_1 = 0.41$, $p = 0.520$; peak SCR: $\chi^2_1 = 0.15$, $p = 0.703$; for Model 3: self-reported valence: $\chi^2_4 = 4.22$, $p = 0.378$; self-reported arousal: $\chi^2_4 = 2.67$, $p = 0.615$; average SCR: $\chi^2_4 = 2.07$, $p = 0.722$; peak SCR: $\chi^2_4 = 0.46$, $p = 0.978$). This indicates that there was neither a main effect of show version, nor an interaction effect of show version with the time-varying effect of intended experience. Intended experience predicted the outcome measures in the show's adapted and original version in the same way. We thus rejected H2.

Differences in lived experience following design interventions

To assess whether the interventions to scene 1 and 7 affected the lived experience of these scenes in particular, we performed two planned comparisons (see Table 2 for the results). H3 states that scene 1 induces more emotional arousal in the show's adapted version than in the original version. Results show that scene 1 in the show's adapted version was reported as slightly more arousing and more positively valenced as compared to the original version. Similar differences were found for average and peak SCR, which were also slightly higher for the show's adapted as compared to the original version. Differences, however, were not significant. We thus found no support for H3.

H4 states that scene 7 induces more emotional arousal in the adapted version of the show as compared to the original version of the show. Scene 7, however, was significantly less arousing in the show's adapted version than in the original version. Compared to the original version of scene 7, the adapted version is also reported as less positively valenced, but the difference was not significant. Average and peak SCR also show a decrease for scene 7 in the show's adapted version as compared to the original version, but these differences were not significant either. We thus rejected H4.

Discussion

This study explored the usefulness of SC and ERMs as tools for evaluating how closely the lived experience of leisure participants follows the intended experience from a leisure provider. We thus compared four lived experience measures of a musical theater show (self-reported valence and arousal, average and peak SCR) against the intended experience from the show's creative development team. Also, we explored whether lived experience measures were affected by employing design interventions to two of the show's scenes (scene 1 and 7), so as to establish whether an attempt of bringing the lived experience more in line with the intended experience was successful.

Results indicate that the intended experience significantly predicted all four lived experience measures. The prediction became stronger as the show progressed over time. In addition, there were no interpretable differences between the adapted and the original versions of scene 1 and 7 across all lived experience measures, and the intended experience did not predict the lived experience measures of the adapted show any differently than those of the original version of the show.

Intended experience significantly predicts lived experience measures

Our data show that both SC and ERMs are significantly predicted from the intended experience. These findings form positive evidence for H1, which states that the time course of lived experience measures is related to the time course of the intended experience. This suggests that in a SE, the experience design can indeed trigger the designed-for emotional responses. Results also show that the correspondence between intended experience and lived experience increases in strength as the musical theater show progresses over time. This effect has not been described in previous literature. We suggest two possible interpretations, together with suggestions to further evaluate the generalizability of this newly found effect to SEs in general.

A first interpretation relates to heterogeneity of the lived experience during the beginning of the show. Although the leisure experience of visiting a musical theater show encompasses more than seeing the show alone, a shortcoming in this study is that the interviews used to establish the intended experience were too much focused on the story of the show itself. As such, the intended experience did not account for emotional responses that are related to non-story aspects. A first type of such non-story related responses might originate in impressions such as being excited for the theater show to begin, being overwhelmed by the overall ambiance of the theater, or feeling energized because of being together hundreds of other people in the audience. Such impressions are likely to be heterogeneous across show visitors and are mostly present during the earlier parts of the show. The fact that the variability in peak SCR seems to be larger during the early scenes of the show (see Figure 4B) supports this interpretation. Once initial excitement is settled and visitors get the grip of the show's story more and more, their experience may become more homogeneous and more in line with the intended experience. The 'initial excitement' hypothesis is supported by the observation that self-reported measures of arousal as well as average and peak SCR for earlier scenes of the show are larger (indicating more excitement) than the intended experience, whereas these differences become smaller for later scenes of the show (see Figures 3 and 4).

Other non-story related emotional responses might result from carry-over effects of experience phases preceding the start of the show. Leisure experiences generally consist of multiple phases, such as anticipation, travel to, on-site, travel back and recollection phases (Clawson & Knetsch, 1966). Previously, it has been suggested that emotions from earlier phases of an experience might carry-over onto the beginning sections of subsequent experience phases (Cojuharenco & Ryvkin, 2008). Varying levels of anticipation, for example, have been demonstrated to significantly affect the emotions in subsequent experience phases (Dixon et al., 2017). Much like initial excitement, we expect preceding experience phases to be highly heterogeneous across participants as well. A carry-over effect could therefore also explain why the lived experience measures in the beginning of the show are subject to greater heterogeneity too, thus having a weaker link with the intended experience than lived experience measures in later sections of the show. Our analysis strategy, however, only focused on the on-site experience from the beginning to the end of the show. Preceding experience phases, such as anticipation, traveling to the theater, parking the car or queueing for the wardrobe, were not included. To establish whether early sections of an experience phase can be subject to a carry-over effect from preceding phases, for future research, we recommend including experience phases that come before the phase of interest as well. Another suggestion would be to record pre-event or resting state measurements before the start of the show, to as to examine participants' initial level of emotional arousal.

A second interpretation, which is not necessarily incompatible with the former, is related to the contents of the musical theater show. First, chances of being picked from the audience as a protagonist for the show is highly unconventional as compared to general theatrical productions. It could be that this unconventionality triggered participants' emotional arousal in addition to the emotions as induced by the story of the show's beginning. Second, all participants in this study were roughly at the same stage of the family life cycle (i.e., being an adult with two children over the age of 6). The show, which is about the carousel of life, addressed this particular family life cycle stage mostly during later scenes. Contents of these later scenes could therefore possibly be experienced as more familiar and more emotional by all participants, leading to a stronger homogeneity of the lived experience across visitors and a better correspondence between lived and intended experience for these scenes in particular. Future research should be aimed at determining whether specific segments of a SE trigger similar emotional responses across specific target audiences based on the recognizability of the SE's contents, and how this compares to other target audiences.

Although the exact correspondence between the temporal dynamics of lived and intended experience has yet to be established more clearly, at the very least, our findings confirm that measures which capture the temporal dynamics of an experience uncover processes that remain invisible to overall evaluation methods. SC and ERMs therefore appear to be useful tools to further advance our understanding of experiences.

Evaluating the effectiveness of design interventions

The design interventions in scene 1 and 7 were intended to increase the induced levels of emotional arousal and in doing so to create a better alignment between the lived experience measures and the intended experience. Against the suggestions of H2, though, lived experience measures of the adapted show were not predicted any better from the intended experience than the show's original version. Also, we reject H3 and H4, which state that the adapted versions of scene 1 and 7 would induce more emotional arousal. For scene 1, we observed higher levels of arousal for the adapted version of the show across all four lived experience measures, but these differences failed to reach significance. This suggests that the intervention to scene 1 changed the lived experience into the desired direction, but that the intervention was not strong enough to produce robust changes. For scene 7, self-reported arousal measures indicated that the adapted version of this scene was emotionally less arousing than the original version. For the three other lived experience measures we also observed lower emotional arousal and lower positive valence levels in the adapted version of scene 7, but differences were non-significant. Together, these findings suggest that the intervention to scene 7 was ineffective at best.

In this case, lived experience measures showed that both of the redesigns did not work any better than the original designs. As the interventions to the show were not theoretically informed but based on creative decisions by the show's creative development, it is difficult to argue why our hypotheses were rejected. As establishing the effectiveness of specific design interventions to a SE is of vital importance for experience designers, we argue that future work should attempt to inform (re)design of show elements based on evidence from quantitative or qualitative research. Arguably, closing the loop between research and creative design is a productive way forward in optimizing experience design. At the very least, our findings show that SC and ERMs are promising research tools for doing so.

Skin conductance versus experience reconstruction measures

In this study, we used both physiological measures (SC) and phenomenological measures (ERMs) of lived experience. Both types of measure were significantly predicted from the intended experience. Nonetheless, the exact relation between the intended experience and the lived experience measures was different between SC and ERMs. Furthermore, the different models showed different effect sizes, with Cohen's f^2 for ERMs prevailing over those for SC. Our observations are in line with the work by Mauss and Robinson (2009), who conclude that physiological and self-reported measures of emotion are associated with unique sources of variance and are thus not necessarily convergent. As such, the two types of measures are not simply interchangeable, nor should one be considered as the gold standard. For future research, we suggest that lived experience measures should be related to experience impact, intent to recommend/revisit, or other relevant experience outcomes, so as to establish which measures have the most explanatory value for which experience outcome. Initial results show that SC significantly predicts differences in proclivity to recommend (Mitas et al., 2020). Comparing SC and ERM measures, Strijbosch et al. (2021) found that intent to recommend is predicted best by reconstructed measures of emotional valence per scene; appreciation by peak SCR measures per scene; overall emotional valence by both average and peak SCR measures per scene and overall emotional arousal by reconstructed measures of emotional arousal per scene. In contrast, Van Bendegom et al. (in press) could not predict overall emotional valence and arousal from SC measures. However, while results seem mixed, empirical work on this matter is only at its dawn. Further research is therefore much needed.

Further considerations

In this study, we selected an experience reconstruction approach to address the phenomenological nature of experience as a complementary method to skin conductance. We used this approach to avoid the methodological issues related to phenomenological approaches that capture the lived experience in real-time, such as experience sampling. Experience sampling bears the disadvantage of disrupting the experience at regular time intervals. In the context of a theater show, this undoubtedly transforms the experience of watching the show. A limitation of experience reconstruction methods, however, is that they employ a retrospective approach to reconstruct measures of lived experience. Lived experience and memory have frequently been found to be different from one another (Kahneman & Riis, 2005), which might affect the validity of reconstructed experience measures. To minimize any experience-memory distortions that would affect the validity of our reconstructed experience measures, we thus kept the time between the experience and the surveying to a strict minimum (i.e., immediately after the show).

Another consideration is that in our participant instructions, we asked participants to relax as much as possible in order to minimize measurement artifacts in particular. In hindsight, this might have affected arousal levels, especially in the earlier parts of the show. Refraining from such instructions is recommended for future research.

A final point of attention is that for theme parks, the biggest portion of their target groups tends to consist of families with children (Anton Clavé, 2007). To ensure that the results of this study would have maximal impact for the theme park hosting the musical theater show, we therefore used a homogenous sample that matched with this particular subset of the target group. Of the participating families, we only subjected the adult family members to our measurement procedures, as recording and storing biometric data from participants under the age of 18 comes with legal constraints. Strictly speaking, this means that the findings in this study can only be generalized to adults in families with children. Especially given the fact that traditional family structures have drastically changed over the last years (OECD, 2011), generalizations to broader theme park audiences should be made with care.

Conclusion

We demonstrate that intended experience predicts the lived experience of a SE as measured with SC and ERMs. The relationship between intended and lived experience increased in strength as the SE unfolded over time. Findings indicate that methods that capture the temporal dynamics of lived experience are useful for evaluating SEs. In addition, SC and ERMs are useful for evaluating SE design interventions, thus constituting a useful feedback tool for SE designers.

Funding

This work was supported by the Center of Expertise Leisure, Tourism & Hospitality (CELTH) under the Storysperience Grant.

ORCID

Wim Strijbosch http://orcid.org/0000-0003-2483-9989
Ondrej Mitas http://orcid.org/0000-0002-7916-0511
John Gelissen http://orcid.org/0000-0002-5232-6649
Marcel Bastiaansen http://orcid.org/0000-0003-2865-7859

References

Anton Clavé, S. (2007). *The global theme park industry*. CABI.

Barnes, S. J., Mattsson, J., & Sørensen, F. (2016). Remembered experiences and revisit intentions: A longitudinal study of safari park visitors. *Tourism Management*, *57*, 286–294. https://doi.org/10.1016/j.tourman.2016.06.014

Bastiaansen, M., Lub, X. D., Mitas, O., Jung, T. H., Passos Ascenção, M., Han, D.-I., Moilanen, T., Smit, B., & Strijbosch, W. (2019). Emotions as core building blocks of an experience. *International Journal of Contemporary Hospitality Management*, *31*(2), 651–668. https://doi.org/10.1108/IJCHM-11-2017-0761

Bastiaansen, M., Oosterholt, M., Mitas, O., Han, D., & Lub, X. (2020). An emotional roller coaster: Electrophysiological evidence of emotional engagement during a roller-coaster ride with virtual reality add-on. *Journal of Hospitality & Tourism Research*, 1–26. https://doi.org/10.1177/1096348020944436

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

Birenboim, A., Dijst, M., Scheepers, F. E., Poelman, M. P., & Helbich, M. (2019). Wearables and location tracking technologies for mental-state sensing in outdoor environments. *The Professional Geographer*, *71*(3), 449–461. https://doi.org/10.1080/00330124.2018.1547978

Del Bosque, I. R., & San Martín, H. (2008). Tourist satisfaction a cognitive-affective model. *Annals of Tourism Research*, *35*(2), 551–573. https://doi.org/10.1016/j.annals.2008.02.006

Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). Springer.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Calvi, L., & Hover, M. (2016). Crossroads: Life changing stories from the second world war. *VIEW Journal of European Television History and Culture*, *5*(10), 55. https://doi.org/10.18146/2213-0969.2016.jethc112

Clawson, M., & Knetsch, J. L. (1966). *Economics of outdoor recreation*. Johns Hopkins Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cojuharenco, I., & Ryvkin, D. (2008). Peak–End rule versus average utility: How utility aggregation affects evaluations of experiences. *Journal of Mathematical Psychology*, *52*(5), 326–335. https://doi.org/10.1016/j.jmp.2008.05.004

Dixon, M. J., Victorino, L., Kwortnik, R. J., & Verma, R. (2017). Surprise, anticipation, and sequence effects in the design of experiential services. *Production and Operations Management*, *26*(5), 945–960. https://doi.org/10.1111/poms.12675

Duerden, M. D., Lundberg, N. R., Ward, P., Taniguchi, S. T., Hill, B., Widmer, M. A., & Zabriskie, R. (2018). From ordinary to extraordinary: A framework of experience types. *Journal of Leisure Research*, *49*(3–5), 196–216. https://doi.org/10.1080/00222216.2018.1528779

Duerden, M. D., Ward, P. J., & Freeman, P. A. (2015). Conceptualizing structured experiences. *Journal of Leisure Research*, *47*(5), 601–620. https://doi.org/10.18666/jlr-2015-v47-i5-6096
Ellis, G. D., Freeman, P. A., Jamal, T., & Jiang, J. (2019). A theory of structured experience. *Annals of Leisure Research*, *22*(1), 97–118. https://doi.org/10.1080/11745398.2017.1312468
Ellis, G. D., Jiang, J., Freeman, P. A., & Lacanienta, A. (2020a). In situ engagement during structured leisure experiences: Conceptualization, measurement, and theory testing. *Leisure Sciences,* 1–19. https://doi.org/10.1080/01490400.2020.1713938
Ellis, G. D., Jiang, J., Freeman, P. A., Lacanienta, A., & Jamal, T. (2020b). Leisure as immediate conscious experience: Foundations, evaluation, and extension of the theory of structured experience. *Journal of Leisure Research*, *51*(5), 581–600. https://doi.org/10.1080/00222216.2020.1754735
Feldman Barrett, L. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23. https://doi.org/10.1093/scan/nsw154
Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, *14*(4), 577–606. https://doi.org/10.1080/026999300402808
Frijda, N. H. (1986). *The emotions. Studies in emotion & social interaction*. Cambridge University Press.
De Geus, S., Richards, G., & Toepoel, V. (2016). Conceptualisation and operationalisation of event and festival experiences: Creation of an event experience scale. *Scandinavian Journal of Hospitality and Tourism*, *16*(3), 274–296. https://doi.org/10.1080/15022250.2015.1101933
Hektner, J. M., Schmidt, J. A., & Csikszentmihaly, M. (2007). *Experience sampling method. Measuring the quality of everyday life*. Sage Publications.
Holland, A. C., & Kensinger, E. A. (2013). Emotion in episodic memory: The effects of emotional contents, emotional state, and motivational goals. In J. Armony & P. Vuilleumier (Eds.), *The Cambridge handbook of human affective neuroscience* (pp. 465–490). Cambridge University Press.
Hosany, S., Martin, D., & Woodside, A. G. (2020). Emotions in tourism: Theoretical designs, measurements, analytics, and interpretations. *Journal of Travel Research*, 1–17. https://doi.org/10.1177/0047287520937079
Jacobs, M. H. (2006). *The production of mindscapes: A comprehensive theory of landscape experience* [dissertation]. Wageningen University.
Jantzen, C. (2013). Experiencing and experiences: A psychological framework. In J. Sundbo & F. Sørensen (Eds.), *Handbook on the experience economy* (pp. 146–170). Edward Elgar Publishing.
Kahneman, D., & Riis, J. (2005). Living and thinking about it: Two perspectives. In F. A. Huppert, N. Baylis, & B. Keverne (Eds.), *The science of well-being*. Oxford University Press.
Kim, J., & Fesenmaier, D. R. (2015). Measuring emotions in real time. *Journal of Travel Research*, *54*(4), 419–429. https://doi.org/10.1177/0047287514550100
Klaus, P., & Maklan, S. (2012). EXQ: A multiple-item scale for assessing service experience. *Journal of Service Management*, *23*(1), 5–33. https://doi.org/10.1108/09564231211208952
Kleiber, D. A., Walker, G. J., & Mannell, R. C. (2011). *A social psychology of leisure*. Venture Publishing.
Lee, H. M., & Smith, S. L. J. (2015). A visitor experience scale: Historic sites and museums. *Journal of China Tourism Research*, *11*(3), 255–277. https://doi.org/10.1080/19388160.2015.1083499
Li, S. (2021). Using self-report and skin conductance measures to evaluate theme park experiences. *Journal of Vacation Marketing*, *27*(2), 133–150. https://doi.org/10.1177/1356766720969738
Li, S., Scott, N., & Walters, G. (2015). Current and potential methods for measuring emotion in tourism experiences: A review. *Current Issues in Tourism*, *18*(9), 805–827. https://doi.org/10.1080/13683500.2014.975679
Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(1), 1–11. https://doi.org/10.1186/s40536-018-0061-2

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Mitas, O., Mitasova, H., Millar, G., Boode, W., Neveu, V., Hover, M., Van den Eijnden, F., & Bastiaansen, M. (2020). More is not better: The emotional dynamics of an excellent experience. *Journal of Hospitality & Tourism Research*, 1–22. https://doi.org/10.1177/1096348020957075

OECD. (2011). *Doing better for families*. OECD Publishing. https://doi.org/10.1787/9789264098732-en

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, *64*(1), 12–40.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295x.110.1.145

Scott, N., Gao, J., & Ma, J. (Eds.). (2017). *Visitor experience design*. CABI.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis. Modeling change and event occurrence*. Oxford University Press.

Smit, B., Melissen, F., Font, X., & Gkritzali, A. (2020). Designing for experiences: a meta-ethnographic synthesis. *Current Issues in Tourism*, 1–19. https://doi.org/10.1080/13683500.2020.1855127

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stienmetz, J., Kim, J., Xiang, Z., & Fesenmaier, D. R. (2021). Managing the structure of tourism experiences: Foundations for tourism design. *Journal of Destination Marketing & Management*, *19*, 100408. https://doi.org/10.1016/j.jdmm.2019.100408

Strijbosch, W., Mitas, O., Van Blaricum, T., Vugts, O., Govers, C., Hover, M., Gelissen, J., & Bastiaansen, M. (2021). When the parts of the sum are greater than the whole: Assessing the peak-and-end-theory for a heterogeneous, multi-episodic tourism experience. *Journal of Destination Marketing & Management*, *20*, 100607–100615. https://doi.org/10.1016/j.jdmm.2021.100607

Strijbosch, W., Mitas, O., Van Gisbergen, M., Doicaru, M., Gelissen, J., & Bastiaansen, M. (2019). From experience to memory: On the robustness of the peak-and-end-rule for complex, heterogeneous experiences. *Frontiers in Psychology*, *10*, 1705. https://doi.org/10.3389/fpsyg.2019.01705

Van Bendegom, C., Mitas, O., Boode, W., De Rooij, P., & Bastiaansen, M. (in press). When the arts are not your cup of tea: Participation frequency and experience in expressive cultural activities. *Journal of Leisure Research*. https://10.1080/00222216.2021.1884820

Zajchowski, C. A. B., Schwab, K. A., & Dustin, D. L. (2017). The experiencing self and the remembering self: Implications for leisure science. *Leisure Sciences*, *39*(6), 561–568. https://doi.org/10.1080/01490400.2016.1209140