Journal of Destination Marketing & Management 20 (2021) 100607

Contents lists available at ScienceDirect

Journal of Destination Marketing & Management

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

When the parts of the sum are greater than the whole: Assessing the peak-and-end-theory for a heterogeneous, multi-episodic tourism experience

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

[a] Academy for Leisure & Events, Breda University of Applied Sciences, Mgr. Hopmansstraat 2, 4817 JS, Breda, the Netherlands
[b] Academy for Tourism, Breda University of Applied Sciences, Mgr. Hopmansstraat 2, 4817 JS, Breda, the Netherlands
[c] Efteling, Europalaan 1, 5171 KW, Kaatsheuvel, the Netherlands
[d] Tilburg School of Social and Behavioral Sciences, Tilburg University, Warandelaan 2, 5037AB, Tilburg, the Netherlands

ARTICLE INFO

Keywords:
Peak-and-end-theory
Experience
Emotions
Physiology
Skin conductance

ABSTRACT

Emotions are a key component of tourism experiences, as emotions make experiences more valued and more memorable. Peak-and-end-theory states that overall experience evaluations are best predicted by the emotions at the most intense and final moments of an experience. Peak-and-end-theory has mostly been studied for relatively simple experiences. Recent insights suggest that peak-and-end-theory does not necessarily hold for tourism experiences, which tend to be more heterogeneous and multi-episodic in nature. Through the novel approach of using electrophysiological measures in combination with experience reconstruction, the applicability of the peak-and-end-theory to the field of tourism is addressed by studying a musical theatre show in a theme park resort. Findings indicate that for a multi-episodic tourism experience, hypotheses from the peak-and-end-theory are rejected for the experience as a whole, but supported for individual episodes within the experience. Furthermore, it is shown that electrophysiology sheds a new light on the temporal dynamics of experience.

1. Introduction

Experience is one of the most important and prevalent concepts in the field of travel and tourism (Scott, Gao, & Ma, 2017). Not only does experience inform definitions of tourism and leisure (Kelly & Godbey, 1992), also, experiences are considered to be the core product of the tourism industries (Mommaas, 2000). Tourism suppliers are therefore fiercely competing on providing tourists with high quality experiences. There are various perspectives on what elements make for an optimal experience, ranging from lived experience elements such as immersion, absorption and engagement (Ellis, Freeman, Jamal, & Jiang, 2017) to overall experience impacts of transformation and meaningfulness (Boswijk, Peelen, & Olthof, 2012; Duerden et al., 2018). Numerous tourism scholars adopt the perspective that experiences are most optimal when they are *memorable*, because only when experiences are memorable they can be remembered and accessed in the future in the first place (Kahneman & Riis, 2005; Zajchowski, Schwab, & Dustin, 2017). Several accounts in the tourism literature have suggested that the memorability of an experience is largely determined by the extent to which emotions are triggered during that experience (Bastiaansen et al., 2019; Del Bosque & San Martín, 2008). This is in line with general psychological theories of episodic memory, which propose that emotions have memory-enhancing properties (Kensinger, 2009; Kensinger & Schacter, 2008). In addition, emotions are known to significantly increase the perceived value of an experience (Bigné, Andreu, & Gnoth, 2005; González-Rodríguez, Domínguez-Quintero, & Paddison, 2019) as well as the intent to revisit and recommend (Lee, 2016). For tourism providers, customer or visitor emotions thus form a crucial ingredient in both the memorability, evaluation and appreciation of their experience products. As such, over the last decade, the study of emotions has become a hot topic in tourism research (Hosany, Martin, & Woodside, 2020; Joo, Cho, Woosnam, & Suess, 2020; Volo, 2021).

In addition to acknowledging the relationship between emotions and the memorability of an experience, it is just as important to understand

* Corresponding author.
E-mail addresses: strijbosch.w@buas.nl (W. Strijbosch), mitas.o@buas.nl (O. Mitas), tvb94@hotmail.com (T. van Blaricum), ovugts@efteling.com (O. Vugts), cagovers@efteling.com (C. Govers), hover.m@buas.nl (M. Hover), j.p.t.m.gelissen@tilburguniversity.edu (J. Gelissen), bastiaansen4.m@buas.nl (M. Bastiaansen).

https://doi.org/10.1016/j.jdmm.2021.100607
Received 11 May 2020; Received in revised form 6 April 2021; Accepted 8 April 2021
Available online 20 April 2021

how these emotions relate to its evaluation and appreciation (Stienmetz, Kim, Xiang, & Fesenmaier, 2021). One influential account for this is offered by the *peak-and-end-theory* (henceforth: PE-theory) (Fredrickson & Kahneman, 1993). The PE-theory states that the evaluation of an experience is determined by the emotions associated with the most intense moment and with the final moment of an experience (Fredrickson & Kahneman, 1993). The PE-theory has proved to be a robust heuristic to close the gap between experience and memory (Fredrickson, 2000). Recent studies, however, suggest that the PE-theory does not necessarily hold for experiences that are emotionally more rich and varied (Cojuharenco & Ryvkin, 2008), or which consist of multiple experiential episodes (Miron-Shatz, 2009; Strijbosch et al., 2019). Arguably, emotional heterogeneity and a multi-episodic nature are some of the core characteristics of tourism experiences (Hammitt, 1980; Mitas, Yarnal, Adams, & Ram, 2012; Nawijn, Mitas, Lin, & Kerstetter, 2012). Work on the PE-theory in a tourism context, however, is rather limited to date (Bastiaansen, Oosterholt, Mitas, Han, & ; Chark, King, & Tang, 2020; Geng, Chen, Lam, & Zheng, 2013; Kemp, Burt, & Furneaux, 2008; Kim & Kim, 2019; Li, 2020; Park, Hahn, Lee, & Jun 2018), which hampers generalizations of the PE-theory to experiences in the field of tourism. Furthermore, most of the work on the PE-theory employs experience sampling methods to measure lived experience (i.e. methods which employ immediate, real-time self-reported measures of lived experience as it is taking place (see e.g. Hektner, Schmidt, & Csikszentmihaly, 2007)), or retrospective approaches to reconstruct experience from memory. This may be problematic, as experience sampling methods disrupt the experience itself, and experience reconstruction methods make use of memory rather than lived experience. Lived experience and memory have frequently been found to be different from one and another (Kahneman & Riis, 2005). Approaches that reconstruct lived experience from memory may thus negatively affect the validity and reliability of the measures used (Larsen & Fredrickson, 1999). This paper tests the PE-theory using state-of-the-art electrophysiological techniques to non-disruptively measure real-time emotions in the lived experience of a highly visited tourism setting: a large musical theatre show in a theme park resort in North-Western Europe that is oriented towards attracting an international audience.

As noted by Scott and Le (2017) in their review on tourism experience, tourism experience is generally approached by two different temporal units of analyses: 1) the trip or vacation as a whole and 2) individual activities that are part of a larger trip or vacation. The present case serves as a representation of the latter category, as the musical theatre show in this study is generally part of a larger trip to the theme park resort. Much like in the present case, grand musical theatre shows and other large scale forms of stage entertainment are important assets of renowned international tourism destinations, such as major theme park areas (Greater Orlando), city districts (West-End in London, Broadway in New York City) or entire cities, such as Las Vegas and Macau (Moss, 2010). The musical theatre show in the present study runs in a theme park resort which attracts an international audience that often stays overnight in the theme park resort's own lodging facilities (i.e. two hotels and two holiday villages). Although the musical theatre show is performed in a theatre located within the boundaries of the theme park, additional tickets are required for attending the show. It is thus offered as a supplementary attraction in addition to the resorts' main attraction (the theme park), targeted towards both visitors that stay at the resort's lodging facilities (i.e. visitors mainly from the Netherlands, Germany, Belgium, France or the United Kingdom) and to visitors who visit the theme park resort in the context of a day-trip. Importantly, as part of the resort destinations' strategy to increase the number of international visitors, the show is completely language free (see section 3.2 for a detailed description of the contents of the show). Therefore, it is argued that this large scale musical theatre show provides a suitable case to study the workings of the PE-theory in the context of an internationally-oriented resort destination. Of course, attending the musical theatre show only provides the experience of a single activity alone in the context of a more complex and encompassing destination experience. In the end, however, the overall destination experience is formed by uniting together the experiences of its individual components (Cetin & Bilgihan, 2016). Understanding the experience of a destination's individual components is therefore just as important as understanding the overall destination experience as a whole. Furthermore, for adequately studying the mechanisms of the PE-theory in a field setting, this limitation is actually a strong point of the present resarch design as it contributes to the ecological validity of the findings concerning this specific element of a destination experience.

2. Theory

2.1. The PE-theory in tourism and leisure experiences

The PE-theory dates back to an experiment conducted by Fredrickson and Kahneman (1993), where they studied the emotional experience of participants in the lab who watched short, plotless movie clips that were either positively or negatively valenced. Participants were asked to continuously indicate their affective experience by moving a slider on a scale that went from "very negative feelings" to "very positive feelings". They found that participants' overall evaluations of the movie clips were predicted best by the most extreme rating given during the clips [peak], as well as the ratings given during the final moments of the clips [end]. Also, a weighted average of peak and end [peak-end] was able to significantly predict overall evaluations of the movie clips. These results led to the formulation of the PE-theory, which in turn incited a cascade of follow-up studies to assess the robustness of the PE-theory for various other experiences. Many of these studies were based on evoking feelings of pain in laboratory settings, such as hearing annoying sounds, submerging one's hand in ice water, or holding one's finger in a closing vise (Ariely, 1998; Ariely and Zauberman, 2000; Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Schreiber & Kahneman, 2000). Clinical experiences have also been extensively studied from a PE-theory perspective (Ariely & Carmon, 2000; Chajut, Caspi, Chen, Hod, & Ariely, 2013; Redelmeier & Kahneman, 1996; Redelmeier, Katz, & Kahneman, 2003). From these studies, the PE-theory has been found to be a robust mechanism in explaining overall evaluations from properties of the lived experience (Fredrickson, 2000).

The tourism literature, too, has a history of studying peaks and ends during experiences (Mitas et al., 2012; Nawijn, 2010; Nawijn, Marchand, Veenhoven, & Vingerhoets, 2010; Nawijn et al., 2012). In these studies, however, peaks and ends have mostly been studied to evaluate the emotional ebb and flow of vacation experiences, without necessarily relating them to overall experience evaluations. It is only recently that the PE-theory, which exactly aims at studying the relationship between experience and overall evaluations, has also been applied to the field of tourism and leisure. Most of the PE-studies on tourism have been conducted in the context of vacations (Chark et al., 2020; Geng et al., 2013; Kemp et al., 2008; Kim & Kim, 2019; Park et al., 2018). Contexts of leisure include watching videos (Baumgartner, Sujan, & Padgett, 1997; Li, Walters, Packer, & Scott, 2019; Müller, Witteman, Spijker, & Alpers, 2019), listening to music (Rozin, Rozin, & Goldberg, 2004; Schäfer, Zimmermann, & Sedlmeier, 2014), playing games (Gutwin, Rooke, Cockburn, Mandryk, & Lafreniere, 2016), riding a roller coaster (Bastiaansen et al., 2020) and engaging with virtual reality (Strijbosch et al., 2019). Yet, although the PE-theory has been corroborated by some of these studies (Baumgartner, Sujan, & Padgett, 1997; Kim & Kim, 2019; Müller et al., 2019; Park et al., 2018; Rozin et al., 2004), just as many studies report that peaks and ends are not always the best predictors of overall evaluations of tourism and leisure experiences (Bastiaansen et al., 2020; Chark et al., 2020; Geng et al., 2013; Gutwin et al., 2016; Kemp et al., 2008; Li, 2020; Li et al., 2019; Schäfer et al., 2014; Strijbosch et al., 2019). How emotions in tourism and leisure experience relate to relevant outcome variables is therefore not necessarily best accounted for by the PE-theory.

We suggest two different yet related explanations for why the PE-theory has had limited predictive value in studies on tourism and leisure experiences. First, the PE-theory has predominantly been confirmed in contexts that yield short and uniform experiences, which cannot straightforwardly be extrapolated to experiences in the context of tourism, which are emotionally richer and more multidimensional in nature (also referred to as heterogeneity). Second, tourism and leisure experiences mostly consist of multiple phases, or episodes. The multi-episodic nature of these experiences has not been taken into account during the development of the PE-theory or in most of its previous assessments, but arguably leads to different evaluation procedures on the one hand, and to even further heterogeneity of the experience on the other. Below, these issues are elaborated on some detail.

2.1.1. The heterogenic and multi-episodic nature of tourism experiences

As mentioned above, the studies that initially led to the formulation of the PE-theory and the ensuing studies that corroborated this theory, have used empirical paradigms that induce relatively short and uniform experiences. Studies on leisure experiences that share these characteristics, such as watching short video clips (Baumgartner et al., 1997; Müller et al., 2019) and listening to music (Rozin et al., 2004), have also found support for the PE-theory. Yet, as most tourism and leisure experiences are characteristic for their emotional heterogeneity (Clawson & Knetsch, 1966; Hammitt, 1980; Stewart, 1998), it is a question whether the PE-theory holds for all experiences in the field of tourism and leisure.

In a meta-analysis on the PE-theory, Cojuharenco and Ryvkin (2008) suggest that for richer and more heterogeneous experiences, emotions at the peak and end moments are often highly correlated with the average of all emotional responses during the experience. In such cases, the average of all emotional responses might therefore be just as good a predictor of the overall evaluation of the experience as traditional peak-and-end-predictors (Cojuharenco & Ryvkin, 2008). In some cases the average of emotional responses is even found to be a better predictor than peaks and ends (Chark et al., 2020; Li, 2020; Li et al., 2019; Miron-Shatz, 2009; Schneider, Stone, Schwartz, & Broderick, 2011; Seta, Haire, & Seta, 2008; Strijbosch et al., 2019).

In addition to the experiential heterogeneity, the multiphasic or multi-episodic nature of most tourism experiences might also limit the applicability of the PE-theory. Arguably, tourism and leisure experiences consist of various experiential episodes (Bastiaansen et al., 2019), that may vary in duration from a chorus of only seconds in an audio track, to scenes of a couple of minutes in a movie or a theatre play, or the individual rides and attractions in a full day visit to a theme park. Ariely and Zauberman, 2000 suggest that "once such [a multi-episode] experience is over, its representation [in memory] no longer contains its pattern but rather only its overall evaluation (…). In addition, we argue that the evaluation of multiple episodes relies on their overall evaluations but not on the hedonic [temporal emotion] profile of those overall intensities." (p. 222). It is suggested, then, that the best predictor for the overall evaluation of the experience as a whole consists of the average of the emotion ratings for all individual episodes (henceforth: *average emotion rating*). This is supported by studies where experience is measured over sequences of aversive sounds (Ariely and Zauberman, 2000), over a virtual reality experience with different scenes (Strijbosch et al., 2019), as well as over entire days filled with activities (Miron-Shatz, 2009). It is also in line with results that suggest average emotion rating as a better predictor than peaks and ends, as mentioned previously. It seems unlikely, though, that all episodes in an experience are given equal weight in computing this average emotion rating. Imagine you visit a pop festival for 12 hours, where in the evening your absolute idol will perform as a headliner for 2 full hours. Does the value of this 2-hour headliner performance then get a similar weight as the other 10 hours of the festival experience, or is the headliner given more weight in computing the average emotion rating for that festival experience? Also, it remains unstudied whether the PE-theory still holds for the individual episodes in an experience, regardless of its limitations for the multi-episode as a whole. Finally, one could argue that with an increase in episodes, heterogeneity will also rise, which on its turn might also affect the workings of the PE-theory.

2.2. Measuring heterogeneous and multi-episodic tourism experiences

2.2.1. Traditional approaches to measuring tourism experiences

As mentioned above, the heterogeneous and multi-episodic nature of tourism experiences begs the question of how the lived experience during each of the different episodes contributes to the overall evaluation of the experience as a whole. Empirically, in the field of tourism and leisure studies this has been approached in three different ways: 1) by manipulating the contents of the experience and measuring the experience outcomes; 2) by retrospectively reconstructing lived experience and 3) by using immediate, real-time measures of lived experience.

The first empirical approach to multi-episodic tourism and leisure experiences entails using an experimental design in which experience is manipulated over different experimental groups (Gutwin et al., 2016; Müller et al., 2019; Park et al., 2018). These studies do not include an ongoing measure of the emotions in an experience, but only measure overall evaluation variables which are then related to the experimental manipulation, rather than to extracted peaks and ends from the ongoing experience measure. Park et al. (2018), for example, manipulated the endings of guided tours. For the control group, the end of the guided tour was a simple bus ride back to the airport, but for the experimental group, this part was changed into a river tour. Müller et al. (2019) also manipulated the final parts of the experience (of a horror movie), and Gutwin et al. (2016) offered various gaming experiences with manipulated peaks and ends. All tourism and leisure studies that use an experimental study design have found that peaks and ends are indeed related to overall experience evaluations. This is in line with other PE-studies that employ an experimental study design outside the domain of tourism and leisure (Ariely, 1998; Ariely & Carmon, 2000; Ariely and Zauberman, 2000; Kahneman et al., 1993; Redelmeier et al., 2003). A limitation of this procedure, though, is that only a limited number of episodes within the experience can be manipulated to be able to interpret the results (in the works mentioned, only peaks or ends have been manipulated), rather than other parameters such as the average emotion rating.

A second empirical approach to multi-episodic tourism and leisure experiences is that of retrospectively reconstructing lived experience through what we term experience reconstruction methods (ERMs). ERMs make use of self-report to reconstruct experiences in two ways: participants are either asked to report on the full experience after it has ended (Chark et al., 2020; Kim & Kim, 2019; Li, 2020; Strijbosch et al., 2019), or participants are asked to report on the aforegoing moments, such as in Geng et al. (2013) and Kemp et al. (2008), where vacationers are asked every evening to complete a brief, retrospective survey on their experience of the whole day. Some of the studies that employ a retrospective approach have found support for the PE-theory in tourism experiences (Geng et al., 2013; Kim & Kim, 2019), but others have challenged the PE-theory through finding non-significant results or finding other predictors that have more explanatory value than those related to the PE-theory (Chark et al., 2020; Geng et al., 2013; Kemp et al., 2008; Li, 2020; Strijbosch et al., 2019). Geng et al. (2013) found support for the PE-theory for evaluations made directly after the experience, but not for evaluations made 3 and 7 weeks after the experience. Results based on retrospectively reconstructing experience are therefore mixed with respect to validating the PE-theory.

Retrospective methods know various limitations. First, a retrospective approach measures *remembered experience* rather than real-time experience. Experience and memory are known to be quite different in nature (see e.g. Kahneman & Riis, 2005) and hence may predict overall evaluations differently. Second, retrospective methods often use single ratings for longer periods of time. For example, participants are asked to give ratings for full days in a week of vacation (Geng et al., 2013; Kemp et al., 2008) that are then representative for this episode within the experience. The resulting ratings, however, may be quite different from those that are based on real-time measures, because of the aggregation within the experiential episodes (Newman, Schwarz, & Stone, 2020). For example, it may be that vacationers experience both positive and negative emotions on a day of vacation, which are then obscured when being asked to give a single rating for the full day. In addition, it can be argued that a day does not necessarily reflect one episode within a vacation, but that the true experiential episodes of a vacation consist of its activities and attractions. As there are several limitations to this empirical approach, it might be useful to complement retrospective data with data collected during the experience itself.

A third empirical approach to multi-episodic tourism and leisure experiences is that of devising immediate, real-time measures of lived experience: experience sampling methods (ESMs). ESMs consist of various approaches, such as a pager that notifies participants when to fill in short, printed self-report forms, or mobile applications that allow for both notifications and digital self-report forms (see Hektner et al. (2007) for an overview). Most of the PE-studies in tourism and leisure employing ESMs do so through the use of a slider on a response device on which participants can continuously report their affective experience, much like in the original Fredrickson and Kahneman study (1993). Experiences studied using this self-reported experience sampling include listening to music and watching videos (Baumgartner et al., 1997; Rozin et al., 2004; Schäfer et al., 2014). All of these studies have found that peaks and ends are indeed related to overall evaluations. It has been noted, though, that consciously providing experience ratings *while the experience is ongoing* may influence the way in which overall evaluations are constructed (Ariely, 1998; Ariely and Zauberman, 2000; Liersch & McKenzie, 2009). More fundamentally, experience sampling might disrupt the experience that you are trying to measure. Although such self-reported experience sampling methods provide a real-time measure of the experience, their generalizability to experience without experience sampling is therefore questionable.

In sum, in order to examine the validity of the PE-theory in the context of heterogeneous and multi-episodic tourism experiences, one must ideally devise measurement techniques that allow for sampling lived experience that may or may not be complemented with retrospective approaches, but that are also able to sample experience in a non-disruptive way. A solution might lie in state-of-the-art methods that have only been recently introduced to the field of tourism: electrophysiological approaches that unobtrusively measure aspects of the experience in real-time with sub-second precision (Bastiaansen et al., 2019; Godovykh & Tasci, 2020; Li, Scott, & Walters, 2015; Steinmetz et al., 2021).

2.2.2. Electrophysiological approaches to measuring tourism experiences

It has been argued previously that electrophysiological measurements are prime candidates for experience sampling with great temporal precision, as they allow for non-disruptively measuring the emotions during lived experience from start to end (Bastiaansen et al., 2019; Li et al., 2015). According to Birenboim, Dijst, Scheepers, Poelman, and Helbich (2019), the use of ambulatory equipment to collect electrophysiological data in the field has various advantages over traditional methods to measure experience. Electrophysiological equipment is able to record real-time physiological signals (such as heart rate, body temperature and skin conductance) that form a more objective measure of emotional experience than self-reported assessments. Furthermore, it is able to measure these physiological signals continuously at high temporal resolution, yielding a detailed measure of emotional experience with sub-second precision. Another advantage is that electrophysiological equipment facilitates investigation of people's physiological signals in the field in a completely non-disruptive way, which offers greater ecological validity as compared to studying physiological signals in a lab. Electrophysiological experience sampling is therefore proposed as a viable alternative to traditional, self-reported experience sampling.

Electrophysiological measurement, however, also carries limitations of its own (Birenboim et al., 2019). Measuring physiology in the field is associated with reduced data quality because of noise and error in the form of motion artifacts. Motion artifacts are high peaks in the recorded signal that result from the movement of sensors relative to the skin, and which do not reflect actual physiology (see Taylor et al. (2015) for a detailed description). Also, there is less control over difficult-to-record environmental factors such as conversations between participants and individual differences in reaction to specific environmental cues, which are crucial for interpreting the results. Data interpretation of such studies is therefore quite demanding. Nonetheless, tourism scholars have successfully started to use such equipment to study emotions in tourism experiences such as city walks (Birenboim et al., 2019; Kim & Fesenmaier, 2015; Osborne & Jones, 2017; Paül i Agustí, Rutllant, & Lasala Fortea, 2019; Shoval, Schvimer, & Tamir, 2018a, 2018b) and museum visits (Kirchberg & Tröndle, 2015; Tröndle, Greenwood, Kirchberg, & Tschacher, 2012), and reviewing tourism marketing materials (Guerrero-Rodríguez, Stepchenkova, & Kirilenko, 2020; Kim, Kim, & Bolls, 2014; Li, Walters, Packer, & Scott, 2018a, 2018b; Li et al., 2019). Thus far, three of these studies have used electrophysiological measures to assess the PE-theory (Bastiaansen et al., 2020; Li, 2020; Li et al., 2019). Within these relatively recent accounts, however, the multi-episodic and heterogeneous nature of the tourism experience remained uncovered.

2.3. The present study

In the present work, the validity of the PE-theory is evaluated for a heterogeneous and multi-episodic tourism experience consisting of a 75-minute musical theatre show in a large internationally-oriented theme park resort in North-Western Europe. A musical theatre show was selected due to its nature as a structured experience (Duerden, Ward, & Freeman, 2015). During the whole experience, people are continuously seated in a theatre while wearing physiological equipment, thus reducing the presence of motion artifacts in the data. Environmental controllability is maintained due to the heavily scripted nature of the theatre show over several clearly delimited scenes, so that all participants are perceiving the same external stimuli at the same well-defined time intervals. A further advantage of the clear division over scenes is that single episodes that together comprise the total experience can easily be identified. This allows for addressing the multi-episodic nature of a highly heterogeneous experience using non-disruptive, real-time physiological measures of affective experience. More specifically, the electrophysiological measure of skin conductance is used to serve as a well-supported proxy for emotional arousal: the extent to which people are emotionally engaged or aroused (Boucsein, 2012). Comparing the approach of electrophysiological experience sampling against the previously used approach of experience reconstruction in earlier studies on the PE-theory, the present paper thus aims to evaluate the robustness of the PE-theory for a heterogeneous and multi-episodic tourism experience.

3. Material and methods

3.1. Participants

Participants were selected from the research panel of the theme park resort. Panel members received an email with information about the study, for which they could sign up. From all subscriptions, 67 participants were selected so as to compose a quota sample that would match with the target group of the show, based on age and family composition. The target group, and hence the population for this study was families with 2 adults and 2 children above the age of 6, as determined by the theme park. Families from both the Netherlands (national visitors) and Germany (international visitors) were then invited to visit the show on a set date in January or February 2019, for which the theme park resort would offer them free tickets. Of the participating families, only the adults were selected to take part in the experiment. They were informed both orally and in writing about the set-up of the study, and subsequently gave their written informed consent in accordance with the declaration of Helsinki.

In total, 67 participants took part in the experiment. Although small samples are common in studies of psychophysiological measures (Li et al., 2019), the number of participants in the present study is substantially larger than the typical $n = 20$–25 as used in traditional laboratory studies on psychophysiology. Of all 67 participants, the physiological data of 10 participants turned out to be either missing due to equipment failure or unusable due to excessive artifacts (see section 3.4.1). These participants were excluded from all further analyses, thus resulting in a final sample of 57 participants. On average, these 57 participants were 40.50 years old ($SD = 8.16$) and consisted of 30 males and 27 females across 33 national (Dutch) and 24 international visitors (Germans).

3.2. Staged experience

The staged experience consisted of a musical theatre show that was performed on a weekly basis in a large-scale theatre that serves as an individual facility within a larger theme park resort. The theme of the show was oriented around the idea of a carousel of life. On average, the show lasted 4470 s ($SD = 36$ s) over 7 performances and consisted of 17 different scenes (for an overview of average scene durations over all performances, see Table 1). Before the beginning of the show, two show hosts pick an (informed) guest out of audience who becomes the protagonist of the show (scene 1). To start the show, the show hosts then magically transform the stage into a big carousel of time, which takes the protagonist back to his childhood to the time where he meets his soon-to-be girlfriend and wife (scene 2). In the following scenes they play children's games with each other (scene 3 and scene 4) and as the carousel of time turns, the protagonist and his female friend grow up (scene 5). The protagonist, now in his puberty, wants to travel the world, whereas his female friend desires to stay home. They separate, and the protagonist sets out on a journey in which he travels across different cultures (scene 6). After having fed on the forbidden fruits of exotic realms, he receives several letters from his female friend back home (scene 7 and scene 8). He realizes he misses her and he returns home where they meet up and start seeing each other (scene 9). After the carousel of time has turned again, they get married (scene 10) and one scene later the woman gives birth to a little baby girl (scene 11). As the carousel of time turns a bit further, they all grow older and end up in the busy grind of parenting and working life (scene 12). At one point, the woman cannot keep up with her husband and child anymore, and she falls to the ground. After a dramatic scene in which she dances with the Death, she eventually dies (scene 13), leaving her husband and child behind. In the next scene, the father and his child work their way out of their sadness through a dance under a magical water curtain (scene 14). As they both pick up their life and as the carousel of time turns again, the now grown-up child gets a baby boy of her own and the protagonist grows into a grandfather. They grow older as a happy family with the protagonist as a content grandfather (scene 15 and scene 16). As the carousel of time turns for a final round, the audience is taken to the childhood period of the baby boy, in which he meets a female friend of his own, thus starting his very own ride on the carousel of life. The show ends with a finale where all actors enter the stage, accompanied by fireworks, bombastic music and applause from the audience (scene 17).

Table 1
Scene durations and per-scene variables.

Scene #	Duration in seconds (*M* (*SD*))	Self-reported valence (*M* (*SD*))	Self-reported arousal (*M* (*SD*))	Average SCR (*M* (*SD*))	Peak SCR (*M* (*SD*))
1	666 (38)	3.456 (0.781)	2.737 (0.955)	0.042 (0.065)	0.299 (0.476)
2	775 (8)	3.912 (0.606)	3.000 (0.824)	0.028 (0.044)	0.374 (0.525)
3	113 (3)	4.088 (0.808)	3.526 (0.908)	0.035 (0.091)	0.158 (0.371)
4	132 (4)	3.386 (0.750)	2.789 (0.750)	0.027 (0.025)	0.157 (0.266)
5	301 (4)	4.491 (0.539)	3.579 (0.981)	0.025 (0.045)	0.223 (0.397)
6	416 (2)	4.000 (0.732)	3.491 (0.848)	0.044 (0.068)	0.244 (0.390)
7	206 (2)	3.895 (0.724)	3.684 (0.736)	0.064 (0.144)	0.467 (0.820)
8	215 (2)	3.737 (0.856)	3.544 (0.803)	0.038 (0.070)	0.259 (0.451)
9	212 (4)	3.421 (0.844)	3.018 (0.954)	0.030 (0.051)	0.217 (0.412)
10	404 (2)	4.316 (0.686)	3.912 (0.830)	0.029 (0.056)	0.232 (0.380)
11	342 (7)	3.719 (0.796)	3.281 (0.861)	0.036 (0.069)	0.205 (0.332)
12	169 (7)	3.281 (0.861)	3.491 (0.848)	0.039 (0.110)	0.227 (0.450)
13	280 (13)	3.333 (1.272)	3.772 (1.035)	0.024 (0.038)	0.208 (0.298)
14	317 (16)	4.088 (1.057)	3.807 (1.043)	0.035 (0.060)	0.302 (0.487)
15	284 (15)	4.000 (0.732)	3.439 (0.866)	0.040 (0.057)	0.292 (0.396)
16	72 (4)	3.737 (0.669)	3.193 (0.743)	0.071 (0.221)	0.221 (0.588)
17	234 (2)	4.368 (0.616)	3.912 (0.714)	0.097 (0.243)	0.472 (0.968)

So as to accommodate for international visitors, the show was almost entirely performed without spoken language. The show heavily drew on mimicry, music and visual spectacle that is understandable for a multinational audience. If any language was used, it was based on a fictional fantasy language mostly spoken by the show hosts.

3.3. Design and procedure

Data were collected during 7 performances of the show over a four-week period in January and February 2019. After having briefed the participants in their own language and having obtained their written informed consent, the experimenter put an Empatica E4 wristband on the wrist of the participants' non-dominant hand (see section 3.4.1). Participants were instructed to sit in the theatre as relaxed as possible, and to not touch or move the wristband during the show. The participants then entered the theatre and watched the theatre show, amidst other, non-participating audience members so as to create an ecologically valid setting. After the show, researchers met participants in a quiet space in the theatre. The experimenter removed the Empatica wristband and participants were given a 60-item post-experience questionnaire in their own language (i.e. Dutch or German), based on validated translations.

Approximately 2 weeks after their visit to the show, all participants received an email with a 5-item questionnaire intended to measure long-term memorability of the show. This questionnaire was completed by 62 of the 67 participants.

3.4. Data collection

3.4.1. Physiological data

Physiological data were recorded with Empatica E4 wearable wristbands (Empatica Inc., USA) that records, amongst others, skin conductance, which is considered to be a reliable index of emotional engagement or arousal (Boucsein, 2012). Skin conductance was continuously sampled at a frequency of 4 Hz and was stored on the Empatica device for further off-line processing. Measurement of physiological responses started at the moment the participants received the Empatica and continued until they were taken off by the experimenter after the show. In order to obtain an indicative time alignment between the physiological recordings and onset/offset of each scene of the show, the experimenter would write down time stamps of the onset and offset of each of the show's performances, as well as around the onset times of the 17 different scenes as described in section 3.2. To do so, the onsets of all 17 scenes were ascribed to clearly observable events in the show. During the show, 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 would occur, the experimenter wrote down the exact time that was simultaneously displayed on the mobile phone.

3.4.2. Post-experience questionnaire

The post-experience questionnaire consisted of two types of questions: questions on per-scene evaluations and questions on overall evaluations of the show. First, participants were asked to evaluate how they had generally felt while watching the show, using two dimensions of emotional valence and emotional arousal adapted from Bradley and Lang (1994). Emotional valence, the extent to which people feel positive or negative, was measured by a 5-point scale ranging from "Very negative" to "Very positive", following the question "Can you indicate to which extent you felt positive or negative during the whole show in general?" Emotional arousal was measured by a 5-point scale ranging from "Very calm" to "Very excited", following the question "Can you indicate to which extent you felt calm or excited during the whole show in general?"

Second, participants were asked to evaluate their emotional valence and arousal for each of the 17 scenes on the show. The scenes were prompted using a picture that was 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. Pictures rather than verbal descriptions were explicitly chosen so as to not prime participants with interpretations that might affect their answering (i.e. labelling scene 13 as "the death scene" might prime participants to evaluate the scene as negatively valenced because of associations with the word "death").

Third, participants were asked to grade the show on a 10-point scale ("How would you grade the show?"), with 1 being a very low grade and 10 being a very high grade. Participant's intent to recommend was also measured using the 11-point scale Net Promotor Score (NPS) (Reichheld, 2003).

3.4.3. Long-term memory questionnaire

Two weeks after their visit to the show, participants received an email with a URL that led to a short web-based questionnaire. This questionnaire contained the two overall evaluation questions from the final part of the post-experience questionnaire, grading the show again on a 10-point scale and indicating their intent to recommend using the 11-point scale NPS.

3.5. Data analysis

3.5.1. Pre-processing of physiological data

The skin conductance data were extracted from the Empatica wristbands, stored on a PC and imported into MATLAB (MathWorks, USA) for further analysis. First, skin conductance data were precisely time-synchronized with the onset and offset of the show, using the time denotations made by the experimenter during the show. Skin conductance segments of 4757 s were then extracted from the recordings, corresponding to the length of the shortest of the 7 performances (performances slightly varied in duration with a range of 103 s (min = 4757 s, max = 4860 s)). Motion artifacts, typically resulting from pressure on the device or from movement of the built-in sensors relative to the skin, were detected and removed using a simple, supervised method for detecting and correcting the skin conductance signal for motion artifacts. Artifacts were detected by applying a z-transform to a moving time window (here 10 s) and visualizing the signal in that time window whenever a z-value exceeded a threshold of ± 3. The experimenter then decided whether or not the detected peak or trough was a motion artifact to be corrected. These decisions were made by comparing the shape of the detected artifact to physiologically plausible skin conductance response shapes (i.e. a sharp rise followed by a gradual decline over multiple seconds (Boucsein, 2012)). In the case of a clear and unambiguous motion artifact, it was removed from the signal by linearly interpolating the signal from the left-hand border of the spike to its right-hand border. In case of ambiguity, the signal was not altered in order to avoid the possibility that true skin conductance responses were removed from the data. The data from 10 participants were excluded from further analysis because their physiological data contained too many artifacts. Of the remaining 57 participants, each participant's skin conductance data were subjected to a continuous deconvolution in order to split the signal into a tonic and a phasic component, using the open-source MATLAB toolbox of Ledalab (Benedek & Kaernbach, 2010). The phasic component or the phasic skin conductance responses (henceforth: SCRs) were then used as a basis for statistical analysis.

Per participant, for each scene, the peak SCR (i.e. the maximum SCR value) as well as the average of the SCR values were calculated for that scene, using the onset and offset times of the 17 different scenes. For most of the scenes, peak and average SCR were based on the skin conductance signal from the entire scene. Some of the scenes, however, would either ask for audience participation (i.e. the audience would stand up and dance in scene 6) or would consistently evoke long-lasting episodes of applause (to accompany a song in scene 8, 11 and 15, and during the finale in scene 17), thus generating excessive motion artifacts during these scenes for all participants across all performances of the show. For these particular scenes, the time segments containing such motion artifacts were left out when calculating peak and average SCR, so as to avoid contamination of peak and average SCR measures (for a visual representation of this procedure, see Fig. 1). Peak and average SCR measures per scene then served as the input for further statistical analysis. Furthermore, from these measures, the following parameters were calculated for the show as a whole (see Table 2 for an overview of the operationalization): *for the average SCR over each scene*: show peak, show end, show peak-end and the mean of the 17 average SCRs of all scenes; *for peak SCR over each scene*: show peak, show end, show peak-end and the mean of the 17 peak SRCs of all scenes.

3.5.2. Pre-processing of per-scene self-report evaluations

The post-experience questionnaire resulted in 17 per-scene emotional valence and arousal ratings (for an overview of the average self-reported valence and arousal per scene, see Table 1), based on which a temporal experience profile of emotional valence and arousal ratings was reconstructed (see Fig. 2A). These profiles served as a representation of the experience as measured through self-report. From these profiles, the following parameters were calculated as an input for the statistical analysis as well (see Table 2 for an overview of the operationalization): *for valence*: peak, trough, end, an average of peak and end [peak-end], an average of trough and end [trough-end] and the average valence ratings over time; *for arousal*: peak, trough, end, peak-end, trough-end and the average arousal ratings over time.

3.5.3. Statistical analyses

In this study, it is examined how various experience predictors relate to overall experience evaluations of the musical theatre show. The statistical analyses follow two different approaches: 1) an approach that considers the show as a non-segmented, single-episode experience in line with traditional studies on the PE-theory (henceforth: the full-show approach) and 2) an approach that considers the show as a multi-episode

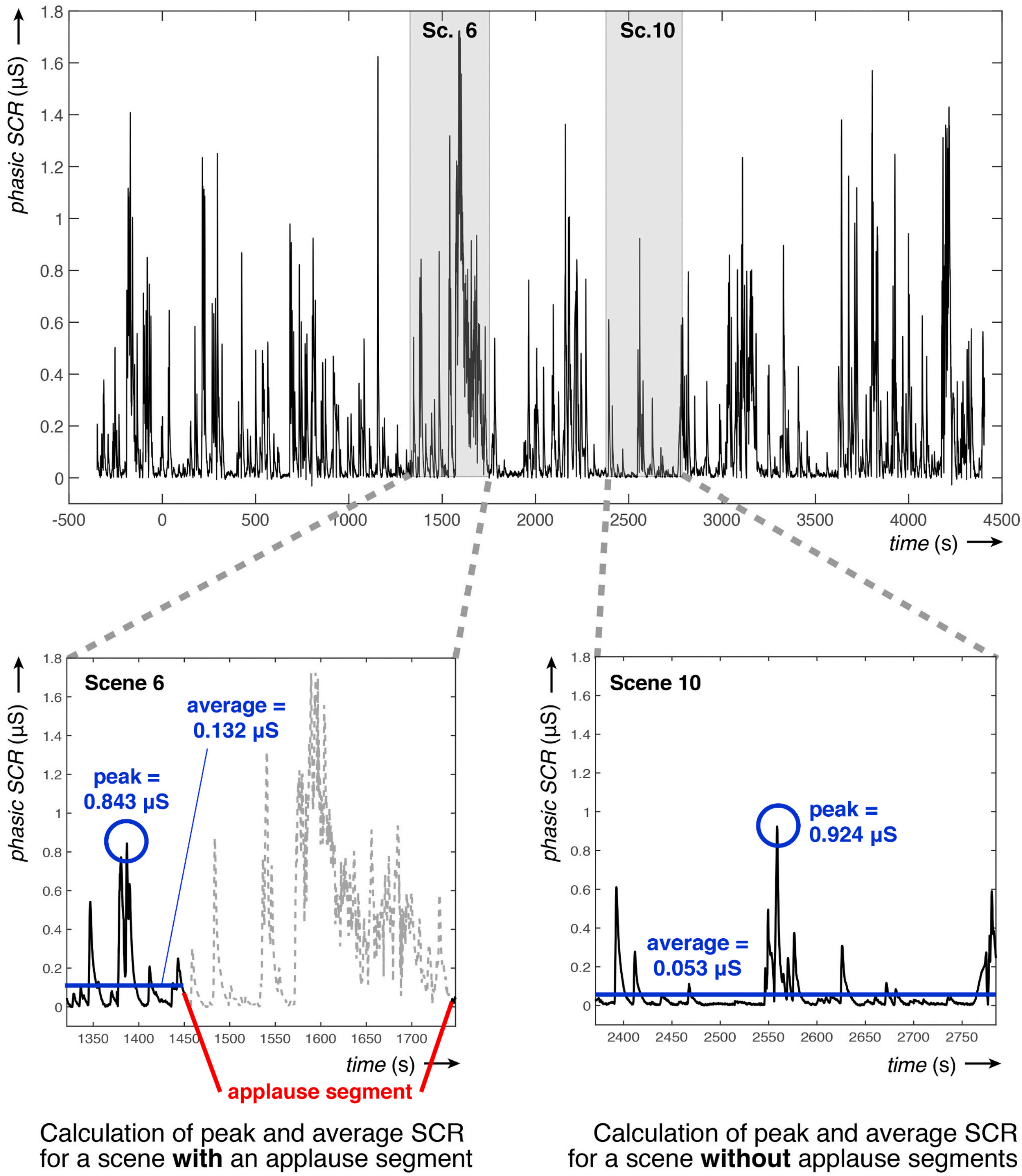

Fig. 1. Calculation of peak and average SCR for scenes with and without applause segments. The top figure displays the phasic component of the full show for one participant. The bottom figures are sections from this phasic component that correspond to the onsets and offsets of scene 6 (bottom left) and scene 10 (bottom right). Scene 10 has no applause segments, and peak and average SCRs are therefore based on the full phasic component of scene 10. Scene 6, however, ends with a segment of applause (marked with the dashed grey line). In calculating peak and average SCRs, the applause segment was left out, and peak and average were calculated from the remaining signal for this scene.

experience, in which the show scenes are considered to be individual episodes that together make up the show as a whole (henceforth: the per-scene approach). In this framework, the multi-episode show thus consists of 17 individual episodes, corresponding to the 17 scenes of the show.

Under the full-show approach, a simple ordinary least squares regression analysis (OLSR) with *one predictor* was performed for any combination of the individual 20 experience predictors (see Table 2) and the 6 outcome variables of overall experience evaluations (overall valence, overall arousal, immediate grade, 2-week-later grade, immediate NPS, 2-week-later NPS). This resulted in 36 OLSRs for valence predictors (6 predictors × 6 outcomes), 36 OLSRs for arousal predictors

Table 2
Operationalization of variables used under the full show approach.

Parameter	Operationalization	M(SD)
Full-show valence ratings		
Peak valence	Most positive valence rating of all 17 scenes	4.825 (0.384)
Trough valence	Most negative valence rating of all 17 scenes	2.526 (0.782)
End valence	Valence rating during final scene	4.368 (0.616
Peak-end valence	Average of peak and end valence	4.596 (0.427)
Trough-end valence	Average of trough and end valence	3.447 (0.450)
Average valence	Average of valence ratings across all 17 scenes	3.837 (0.333)
Full-show arousal ratings		
Peak arousal	Most intense arousal rating of all 17 scenes	4.632 (0.522)
Trough arousal	Most calm arousal rating of all 17 scenes	2.053 (0.692)
End arousal	Arousal rating during final scene	3.912 (0.714)
Peak-end arousal	Average of peak and end arousal	4.272 (0.527)
Trough-end arousal	Average of trough and end arousal	2.982 (0.481)
Average arousal	Average of arousal ratings across all 17 scenes	3.422 (0.461)
Full-show average SCR values		
Peak average SCR	Highest average SCR value of all 17 scenes	0.135 (0.248)
End average SCR	Average SCR value during final scene	0.097 (0.243)
Peak-end average SCR	Average of peak and end average SCR	0.116 (0.243)
Average of average SCR	Average of average SCR values across all 17 scenes	0.041 (0.077)
Full-show peak SCR values		
Peak peak SCR	Highest peak SCR value of all 17 scenes	0.820 (1.083)
End peak SCR	Peak SCR value during final scene	0.472 (0.968)
Peak-end peak SCR	Average of peak and end peak SCR	0.646 (1.001)
Average of peak SCR	Average of peak SCR values across all 17 scenes	0.268 (0.392)

(6 predictors × 6 outcomes), 24 OLSRs for average SCR predictors (4 predictors × 6 outcomes) and 24 OLSRs for peak SCR predictors (4 predictors × 6 outcomes). To reduce the family-wise error rate, a Bonferroni correction was applied, yielding criteria for significance of $\alpha_{FW} = 0.008$ for the family of self-reported valence and arousal predictors and $\alpha_{FW} = 0.013$ for the family of average and peak SCR predictors (based on 6 and 4 OLSRs in one family of tests, respectively). In order to compare the predictive value of the different predictors, for each regression, the significance (*p*-values) of the *F*-statistic and the coefficient of determination (R^2-values) per regression analysis are reported.

Under the per-scene approach, a multiple OLSR was performed in which the experience predictors from all 17 scenes were included as *17 predictors* within one regression model. The regression models were based on any combination of the 4 predictors suited for these analyses (self-reported valence per scene, self-reported arousal per scene, average SCR per scene and peak SCR per scene) with the 6 outcome variables of overall experience evaluations (see above). This resulted in 6 OLSRs for valence predictors per scene (4 predictor types × 6 outcomes), 6 OLSRs for arousal predictors per scene (4 predictor types × 6 outcomes), 6 OLSRs for average SCR per scene (4 predictor types × 6 outcomes) and 6 OLSRs for peak SCR per scene (4 predictor types × 6 outcomes). As the regression models were not part of the same family of tests, in the per-scene approach, a Bonferroni correction was deemed unnecessary. In order to compare the predictive value of the different type of per-scene parameters, for each regression model the significance (*p*-values) of the *F*-statistic are reported. Given the large number of predictors in each model (i.e. 17), adjusted R^2-values (R^2_{adj}) are reported as compared to regular R^2-values, as this statistic adjusts R^2 for the number of explanatory terms in a regression model (Theil, 1961). If the regression model proved to be significant, the number of predictors in the model with a regression coefficient that is significantly different from 0 are also reported.

4. Results

The grand average temporal profile of self-reported valence and arousal over all participants can be found in Fig. 2A. For the per-scene ratings, valence ranged from 1 to 5 and the average of all ratings across segments and participants indicated an experience of neutral to positive valence ($M = 3.810$; $SD = 0.331$). Per-scene arousal ratings ranged from 1 to 5 as well and the average across scenes and participants indicated a moderate level of arousal ($M = 3.34$; $SD = 0.484$). The grand average temporal profile of SCR is presented in Fig. 2B.

Immediately after the performance, respondents indicated that, in general, they experienced the show as highly positively valenced ($M = 4.54$; $SD = 0.538$) and moderately arousing ($M = 3.21$; $SD = 0.868$). Immediately after the performance, respondents evaluated the show with a high grade ($M = 8.70$; $SD = 1.085$) and evaluated the show with an equally high grade two weeks later ($M = 8.69$; $SD = 0.928$, $t_{53} = 0.594$; $p = 0.617$). Also, based on the NPS-item, respondents indicated that they were very likely to recommend the show to colleagues or acquaintances, giving an average of 8.51 ($SD = 1.649$) directly after the performance and an average of 8.37 after 2 weeks ($SD = 1.391$), with the two occasions again not differing from each other ($t_{53} = 1.990$; $p = 0.052$).

4.1. Analyses under the full-show approach

Results from the regression analyses in which overall experience evaluations were predicted from full-show self-reported valence and arousal are presented in Tables 3 and 4. Findings show that valence predictors yield more statistically significant regression models (12 in total) than arousal predictors (3 in total). For valence predictors, the predictor with the highest explanatory value is not consistent over the various overall evaluation variables. Overall valence and 2-wk-later NPS are predicted best by peak valence, whereas immediate grade is predicted equally well by peak-end and average valence, and immediate NPS is predicted best by average valence. Trough valence and trough-end valence did not significantly predict any of the overall evaluation variables. For the arousal predictors, average arousal consistently leads to the highest portions of explained variance amongst three of overall evaluation variables (overall arousal, immediate grade and immediate NPS). Overall valence, 2-wk-later grade and 2-wk-later NPS are not significantly predicted from arousal predictors at all. All other arousal predictors (peak, trough, end, peak-end and trough-end) did not yield any significant regression models.

Results from the regression analyses in which overall experience evaluations were predicted from full-show average SCR and full-show peak SCR are presented in Tables 5 and 6. The analyses indicate that none of the outcome variables (overall valence, overall arousal, immediate grade and NPS, and 2-week-later grade and NPS) could be significantly predicted from full-show average SCR and full-show peak SCR values.

4.2. Analyses under the per-scene framework

Results from the regression analyses in which overall experience

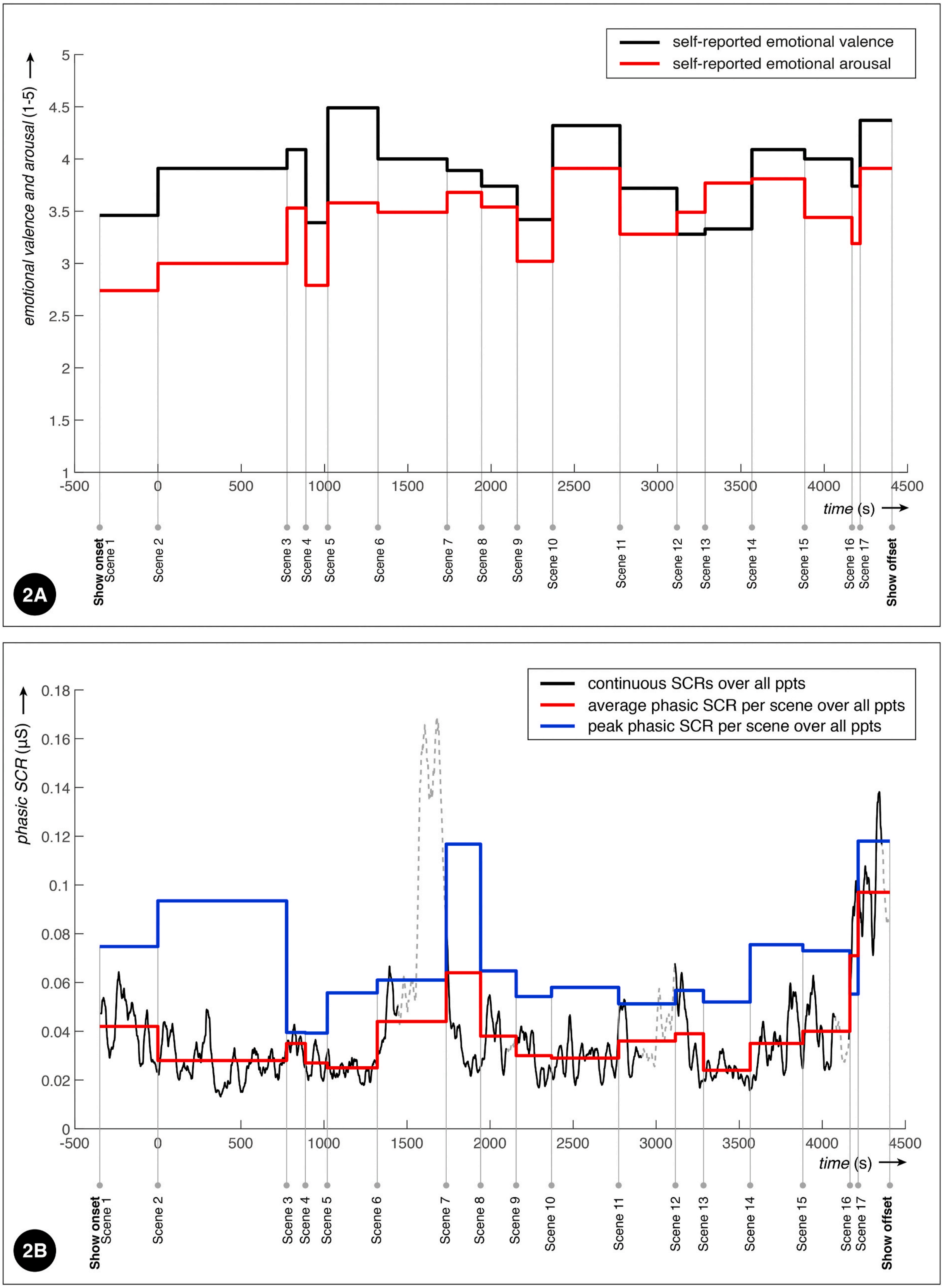

Fig. 2. 2A (top): The grand average of the self-reported emotional valence and arousal profiles over all participants. Emotional valence is indicated in black, emotional arousal is indicated in red. 2B (bottom): The grand average of the phasic component over all participants, as well as the grand average of the average SCRs per scene and the peak SCRs per scene. The sections that are not included in the calculation of the average and peak SCRs are indicated with a grey, dashed line in the bottom figure.

Table 3
Results of full-show regression analyses for self-reported valence predictors.

Predictor	Overall valence	Overall arousal	Immediate grade	Immediate NPS	2-wk-later grade	2-wk-later NPS
Peak	$R^2 = 0.219$ $F = 15.172$ $p < 0.001$	$R^2 = 0.004$ $F = 0.208$ $p = 0.650$	$R^2 = 0.150$ $F = 9.674$ $p = 0.003$	$R^2 = 0.136$ $F = 8.960$ $p = 0.005$	$R^2 = 0.014$ $F = 0.723$ $p = 0.399$	$R^2 = 0.167$ $F = 10.420$ $p = 0.002$
Trough	$R^2 = 0.005$ $F = 0.287$ $p = 0.594$	$R^2 = 0.007$ $F = 0.364$ $p = 0.549$	$R^2 = 0.000$ $F = 0.000$ $p = 0.993$	$R^2 = 0.004$ $F = 0.238$ $p = 0.628$	$R^2 = 0.025$ $F = 1.335$ $p = 0.253$	$R^2 = 0.012$ $F = 0.622$ $p = 0.434$
End	$R^2 = 0.084$ $F = 4.958$ $p = 0.030$	$R^2 = 0.006$ $F = 0.328$ $p = 0.569$	$R^2 = 0.189$ $F = 12.790$ $p = 0.001$	$R^2 = 0.072$ $F = 4.297$ $p = 0.043$	$R^2 = 0.007$ $F = 0.375$ $p = 0.543$	$R^2 = 0.070$ $F = 3.907$ $p = 0.053$
Peak-end	$R^2 = 0.177$ $F = 11.600$ $p = 0.001$	$R^2 = 0.001$ $F = 0.042$ $p = 0.838$	$R^2 = 0.237$ $F = 17.098$ $p < 0.001$	$R^2 = 0.130$ $F = 8.192$ $p = 0.006$	$R^2 = 0.013$ $F = 0.670$ $p = 0.417$	$R^2 = 0.138$ $F = 8.290$ $p = 0.006$
Trough-end	$R^2 = 0.082$ $F = 4.834$ $p = 0.032$	$R^2 = 0.017$ $F = 0.960$ $p = 0.332$	$R^2 = 0.088$ $F = 5.297$ $p = 0.025$	$R^2 = 0.058$ $F = 3.400$ $p = 0.071$	$R^2 = 0.006$ $F = 0.329$ $p = 0.569$	$R^2 = 0.007$ $F = 0.393$ $p = 0.534$
Average	$R^2 = 0.189$ $F = 12.590$ $p = 0.001$	$R^2 = 0.039$ $F = 2.199$ $p = 0.144$	$R^2 = 0.233$ $F = 16.722$ $p = 0.001$	$R^2 = 0.209$ $F = 14.490$ $p < 0.001$	$R^2 = 0.034$ $F = 1.812$ $p = 0.184$	$R^2 = 0.120$ $F = 7.113$ $p = 0.010$

Note. α is set at 0.008 due to a Bonferroni correction. Significant regression models are marked in grey.

Table 4
Results of full-show regression analyses for self-reported arousal predictors.

Predictor	Overall valence	Overall arousal	Immediate grade	Immediate NPS	2-wk-later grade	2-wk-later NPS
Peak	$R^2 = 0.075$ $F = 4.377$ $p = 0.041$	$R^2 = 0.020$ $F = 1.077$ $p = 0.304$	$R^2 = 0.045$ $F = 2.598$ $p = 0.113$	$R^2 = 0.049$ $F = 2.841$ $p = 0.098$	$R^2 = 0.039$ $F = 2.085$ $p = 0.155$	$R^2 = 0.083$ $F = 4.703$ $p = 0.035$
Trough	$R^2 = 0.000$ $F = 0.020$ $p = 0.889$	$R^2 = 0.026$ $F = 1.431$ $p = 0.237$	$R^2 = 0.014$ $F = 0.755$ $p = 0.389$	$R^2 = 0.003$ $F = 0.163$ $p = 0.688$	$R^2 = 0.000$ $F = 0.000$ $p = 0.991$	$R^2 = 0.001$ $F = 0.073$ $p = 0.788$
End	$R^2 = 0.048$ $F = 2.730$ $p = 0.104$	$R^2 = 0.086$ $F = 5.065$ $p = 0.029$	$R^2 = 0.097$ $F = 5.902$ $p = 0.018$	$R^2 = 0.079$ $F = 4.723$ $p = 0.034$	$R^2 = 0.002$ $F = 0.125$ $p = 0.725$	$R^2 = 0.046$ $F = 2.529$ $p = 0.118$
Peak-end	$R^2 = 0.081$ $F = 4.748$ $p = 0.034$	$R^2 = 0.072$ $F = 4.176$ $p = 0.046$	$R^2 = 0.100$ $F = 6.114$ $p = 0.017$	$R^2 = 0.090$ $F = 5.459$ $p = 0.023$	$R^2 = 0.017$ $F = 0.894$ $p = 0.349$	$R^2 = 0.083$ $F = 4.737$ $p = 0.034$
Trough-end	$R^2 = 0.031$ $F = 1.733$ $p = 0.194$	$R^2 = 0.111$ $F = 6.718$ $p = 0.012$	$R^2 = 0.099$ $F = 6.038$ $p = 0.017$	$R^2 = 0.061$ $F = 3.593$ $p = 0.063$	$R^2 = 0.001$ $F = 0.066$ $p = 0.798$	$R^2 = 0.036$ $F = 1.918$ $p = 0.172$
Average	$R^2 = 0.077$ $F = 4.473$ $p = 0.039$	$R^2 = 0.293$ $F = 22.390$ $p < 0.001$	$R^2 = 0.176$ $F = 11.774$ $p = 0.001$	$R^2 = 0.135$ $F = 8.584$ $p = 0.005$	$R^2 = 0.049$ $F = 2.687$ $p = 0.107$	$R^2 = 0.093$ $F = 5.313$ $p = 0.025$

Note. α is set at 0.008 due to a Bonferroni correction. Significant regression models are marked in grey.

evaluations were predicted from self-reported per-scene valence and arousal values are presented in Table 7. Findings show that overall arousal, immediate grade and immediate NPS can be significantly predicted from self-reported arousal per scene, self-reported valence per scene, or both. Self-reported arousal per scene, then, yields regression models with higher portions of explained variance as compared to self-reported valence per scene. In addition, findings show that the portions of explained variance from both self-reported valence and arousal are higher under the per-scene approach than under the full-show approach (for details, see Table 7 as compared to Tables 3 and 4 Overall valence, 2-wk-later grade and 2-wk-later NPS, then, could not be predicted from per-scene valence and arousal ratings.

Results from the regression analyses in which overall experience evaluations were predicted from per-scene average and peak SCR values

Table 5
Results of full-show regression analyses for average SCR predictors.

Predictor	Overall valence	Overall arousal	Immediate grade	Immediate NPS	2-wk-later grade	2-wk-later NPS
Peak	$R^2 = 0.005$ $F = 0.250$ $p = 0.619$	$R^2 = 0.002$ $F = 0.119$ $p = 0.731$	$R^2 = 0.000$ $F = 0.023$ $p = 0.880$	$R^2 = 0.000$ $F = 0.000$ $p = 0.987$	$R^2 = 0.010$ $F = 0.515$ $p = 0.476$	$R^2 = 0.004$ $F = 0.186$ $p = 0.668$
End	$R^2 = 0.001$ $F = 0.042$ $p = 0.839$	$R^2 = 0.000$ $F = 0.005$ $p = 0.943$	$R^2 = 0.000$ $F = 0.011$ $p = 0.915$	$R^2 = 0.003$ $F = 0.153$ $p = 0.697$	$R^2 = 0.001$ $F = 0.078$ $p = 0.782$	$R^2 = 0.020$ $F = 1.054$ $p = 0.309$
Peak-end	$R^2 = 0.002$ $F = 0.127$ $p = 0.723$	$R^2 = 0.001$ $F = 0.045$ $p = 0.833$	$R^2 = 0.000$ $F = 0.001$ $p = 0.981$	$R^2 = 0.001$ $F = 0.041$ $p = 0.839$	$R^2 = 0.005$ $F = 0.254$ $p = 0.616$	$R^2 = 0.010$ $F = 0.535$ $p = 0.468$
Average	$R^2 = 0.018$ $F = 1.016$ $p = 0.318$	$R^2 = 0.003$ $F = 0.131$ $p = 0.719$	$R^2 = 0.000$ $F = 0.007$ $p = 0.935$	$R^2 = 0.004$ $F = 0.207$ $p = 0.651$	$R^2 = 0.002$ $F = 0.108$ $p = 0.744$	$R^2 = 0.001$ $F = 0.028$ $p = 0.868$

Note. α is set at 0.013 due to a Bonferroni correction.

Table 6
Results of full-show regression analyses for peak SCR predictors.

Predictor	Overall valence	Overall arousal	Immediate grade	Immediate NPS	2-wk-later grade	2-wk-later NPS
Peak	$R^2 = 0.000$ $F = 0.002$ $p = 0.968$	$R^2 = 0.002$ $F = 0.113$ $p = 0.738$	$R^2 = 0.000$ $F = 0.011$ $p = 0.917$	$R^2 = 0.000$ $F = 0.008$ $p = 0.929$	$R^2 = 0.005$ $F = 0.236$ $p = 0.629$	$R^2 = 0.004$ $F = 0.228$ $p = 0.635$
End	$R^2 = 0004$ $F = 0.196$ $p = 0.660$	$R^2 = 0.005$ $F = 0.273$ $p = 0.603$	$R^2 = 0.006$ $F = 0.349$ $p = 0.557$	$R^2 = 0.015$ $F = 0.826$ $p = 0.367$	$R^2 = 0.001$ $F = 0.042$ $p = 0.838$	$R^2 = 0.036$ $F = 1.963$ $p = 0.167$
Peak-end	$R^2 = 0.001$ $F = 0.037$ $p = 0.849$	$R^2 = 0.003$ $F = 0.189$ $p = 0.666$	$R^2 = 0.002$ $F = 0.117$ $p = 0.734$	$R^2 = 0.004$ $F = 0.235$ $p = 0.629$	$R^2 = 0.001$ $F = 0.027$ $p = 0.870$	$R^2 = 0.016$ $F = 0.863$ $p = 0.357$
Average	$R^2 = 0.017$ $F = 0.908$ $p = 0.345$	$R^2 = 0.002$ $F = 0.094$ $p = 0.760$	$R^2 = 0.001$ $F = 0.028$ $p = 0.868$	$R^2 = 0.010$ $F = 0.530$ $p = 0.470$	$R^2 = 0.004$ $F = 0.195$ $p = 0.661$	$R^2 = 0.002$ $F = 0.110$ $p = 0.742$

Note. α is set at 0.013 due to a Bonferroni correction.

Table 7
Results of per-scene regression analyses for both self-reported and physiological predictors.

Predictor	Overall valence	Overall arousal	Immediate grade	Immediate NPS	2-wk-later grade	2-wk-later NPS
Self-reported valence per scene	$R^2_{adj} = 0.144$ $F = 1.544$ $p = 0.131$ $n_{predictors} = 0$	$R^2_{adj} = 0.096$ $F = 1.343$ $p = 0.220$ $n_{predictors} = 0$	$R^2_{adj} = 0.275$ $F = 2.249$ $p = 0.018$ $n_{predictors} = 1$	$R^2_{adj} = 0.289$ $F = 2.337$ $p = 0.014$ $n_{predictors} = 3$	$R^2_{adj} = -0.013$ $F = 0.961$ $p = 0.517$ $n_{predictors} = 0$	$R^2_{adj} = 0.188$ $F = 1.722$ $p = 0.084$ $n_{predictors} = 0$
Self-reported arousal per scene	$R^2_{adj} = 0.042$ $F = 1.142$ $p = 0.354$ $n_{predictors} = 0$	$R^2_{adj} = 0.447$ $F = 3.618$ $p < 0.001$ $n_{predictors} = 2$	$R^2_{adj} = 0.176$ $F = 1.702$ $p = 0.085$ $n_{predictors} = 0$	$R^2_{adj} = 0.231$ $F = 1.990$ $p = 0.038$ $n_{predictors} = 1$	$R^2_{adj} = -0.205$ $F = 0.496$ $p = 0.951$ $n_{predictors} = 0$	$R^2_{adj} = -0.118$ $F = 0.670$ $p = 0.810$ $n_{predictors} = 0$
Average SCR per scene	$R^2_{adj} = 0.292$ $F = 2.334$ $p = 0.015$ $n_{predictors} = 5$	$R^2_{adj} = -0.119$ $F = 0.657$ $p = 0.823$ $n_{predictors} = 0$	$R^2_{adj} = 0.213$ $F = 1.894$ $p = 0.050$ $n_{predictors} = 2$	$R^2_{adj} = 0.097$ $F = 1.355$ $p = 0.212$ $n_{predictors} = 0$	$R^2_{adj} = 0.156$ $F = 1.577$ $p = 0.123$ $n_{predictors} = 0$	$R^2_{adj} = 0.192$ $F = 1.742$ $p = 0.080$ $n_{predictors} = 0$
Peak SCR per scene	$R^2_{adj} = 0.294$ $F = 2.346$ $p = 0.014$ $n_{predictors} = 6$	$R^2_{adj} = -0.088$ $F = 0.738$ $p = 0.746$ $n_{predictors} = 0$	$R^2_{adj} = 0.321$ $F = 2.554$ $p = 0.008$ $n_{predictors} = 8$	$R^2_{adj} = 0.148$ $F = 1.571$ $p = 0.121$ $n_{predictors} = 0$	$R^2_{adj} = 0.010$ $F = 1.032$ $p = 0.450$ $n_{predictors} = 0$	$R^2_{adj} = 0.125$ $F = 1.444$ $p = 0.173$ $n_{predictors} = 0$

Note. Significant regression models are marked in grey. $n_{predictors}$ denotes the number of significant predictors in the regression model.

are presented in Table 7 as well. Findings show that both overall valence and immediate grade are significantly predicted from average SCR per scene as well as peak SCR per scene. Both for overall valence and the immediate grade, peak SCR per scene consistently yields the highest portions of explained variance. Much like self-reported values per scene, the portions of explained variance from both average SCR and peak SCR under the per-scene approach are higher than those under the full-show approach. Also, per scene models using SCR predictors show higher percentages of explained variance as compared to per scene valence and arousal reports. Finally, 2-wk-later grade and 2-wk-later NPS could not be predicted from per-scene average SCR and peak SCR values.

5. Discussion

This study assessed the applicability of the PE-theory to relating immediate experience to remembered experience during a musical theatre show offered in an internationally-oriented theme park resort. More specifically, the robustness of the PE-theory was adressed using state-of-the-art electrophysiological equipment that is able to non-disruptively measure lived experience, and was compared against traditional experience reconstruction methods. In addition, the validity of the PE-theory was explored in the context of a heterogeneous and multi-episodic tourism experience by not only applying the PE-theory to the experience as a whole, but also to the individual episodes that it consists of. Results indicate that multi-episodic models using per-scene predictions yield higher portions of shared variance with overall experience evaluations than models which consider the full show as a whole. This is consistent across both self-reported and physiological measures of lived experience. In addition, while predictors related to the PE-theory are not always the best under the full-show approach, in the multi-episode approach, peak does prove to be the predictor that yields the highest portion of shared variance with overall experience evaluations. These results shed a new light on the workings of the PE-theory for heterogeneous, multi-episode experiences, such as those in the field of tourism.

5.1. Per-episode predictions work better than full-experience predictions

Previously used approaches to multi-episode experience evaluation (in the present study: the full-show approach) extract measures of peaks, troughs, ends and averages from start to end of the full experience. In the present study, this approach yields significant results, with all overall experience evaluations (except for the grade as provided two weeks later) being significantly predicted from peak, end, peak-end and average ratings for valence predictors, and average ratings for arousal predictors. This observation is thus in line with the PE-theory, as also found in several other studies that employed this approach (Ariely and Zauberman, 2000; Miron-Shatz, 2009; Strijbosch et al., 2019). In addition to the full-show approach, a per-scene approach was also used. In this approach, the fact that each individual episode in an experience may differently contribute to overall experience evaluations was accounted for. Under this approach as well, almost all overall experience evaluations (except for the grade and NPS as provided two weeks later) were significantly predicted from self-reported valence, self-reported arousal, average SCR and peak SCR. More importantly, though, the per-scene approach consistently accounts for *more* variance in the outcome variables than the full-show approach. This suggests that in the case of a heterogeneous, multi-episode tourism experience, per-episode predictions work better than full-experience predictions.

This has several implications. First, the findings illustrate that not all episodes within an experience equally contribute to overall experience evaluations. At most, only 8 of the 17 scenes within the theatre show significantly contributed to predicting outcome measures (for the immediate grade being predicted from peak SCR per scene, $n_{\text{predictors}} = 8$, see Table 7). In addition, the findings indicate that higher emotional arousal does not always correspond with more positive overall evaluations of the show. Some of the significant regression coefficients show a positive correlation between emotional arousal in a scene and overall experience evaluations. Some of the regression coefficients, however, also demonstrate a negative correlation between the two. Thus, an increase in emotional arousal during some of the scenes has a positive effect on the overall evaluation, yet for other scenes less emotional arousal would be better. This is in line with recent work in the tourism literature, which suggests that more emotion is not always better (see e.g. Mitas et al., 2020; Nawijn & Fricke, 2015).

In sum, the findings imply that the temporal dynamics of a lived tourism experience carry more information than peaks, ends and averages of the overall experience, and that segmentation of a tourism experience into experiential episodes is a worthwile endeavour. This is in line with recent conceptualizations of tourism and leisure experiences as multi-episodic (Bastiaansen et al., 2019; Steinmetz et al., 2021). Traditionally, tourism experiences are mostly segmented into pre-, direct and post-exposure phases (see Godovykh and Tasci (2020) for a recent review). In addition to this broad division into experience phases, the current findings imply that it may be beneficial to incorporate even more detailed and fine-grained per-episode time information within these phases of tourism experiences as well.

5.2. The PE-theory for full experiences versus individual episodes

Another finding is that predictions from average valence and arousal perform equally or even better than peaks and ends under the full-show approach, but that peak SCRs consistently outperform average SCRs under the per-scene approach. For individual episodes, the findings are thus in line with the PE-theory, but not for the full-show approach. This is in line with suggestions from Ariely and Zauberman (2000) and Miron-Shatz (2009) that the evaluation of individual episodes follows the PE-theory, but that the overall evaluations of multi-episode experiences follow other heuristics that connect lived experience to overall evaluations.

5.3. Physiological measures of emotion only work for individual episodes

As said, the continuous recording that is inherent to physiological measurement opens up new paths of studying the temporal dynamics of experience. In this study, per-episode physiological measures of emotion yield the highest portions of explained variance when predicting overall experience evaluations. In contrast, none of the physiological measures of emotion could predict overall evaluations when peaks, ends and average emotion ratings were computed for the entire experience (see Tables 5 and 6), which is in accordance with earlier electrophysiological PE-studies in the tourism literature (Bastiaansen et al., 2020; Li, 2020). Under the full-experience framework, self-reported valence and arousal do significantly predict overall experience evaluations. The abovementioned results are in line with Ariely and Zauberman's (2000) suggestion that evaluations of multi-episode experiences rely more on the *evaluation* of individual episodes than on the temporal pattern of emotions within these episodes. This view apparently contrasts with findings from Li et al. (2019), who report significant relationships between peak, end and average SCR and experience outcomes for tourism advertising videos. However, the advertising videos used in their study are rather short (60–90 s), and therefore can be viewed as individual episodes, rather than as a multi-episode experience. Exactly how multi-episodic natures of an experience affect the workings of the PE-theory for physiological data remains to be studied more systematically, but the present data show at the very least that for per-episode approaches physiological data yield higher portions of explained variance for such overall experience evaluations as overall valence and the grade immediately provided after the show. As suggested previously, physiological measures of emotion are thus a promising way to study experiences in tourism and leisure (Bastiaansen et al., 2019; Godovykh & Tasci, 2020; Li et al., 2015).

5.4. Further considerations

In this study, a heterogeneous, multi-episodic tourism experience was studied that, due to its staged and theatrical nature, was easy to dissect into individual episodes, following the scene division from the experience provider. The reason for doing so is that in this way experiential episodes are similar across participants, which reduces between-subject differences that otherwise form a source of error to the necessary within-subject analyses. A provider-based segmentation is not new to the field of tourism research, in which experiences have long been dissected into "stages" or "phases" from an experience provider point of view (see e.g. Clawson & Knetsch, 1966). Experience, however, is highly personal, based on individual memories and frames of reference, and it is thought that people segment their experiences according to their own mental models of the world (Bastiaansen et al., 2019). For destination management, too, it is suggested that in studies on experience the demand-side should be taken more into account thant the supply-side alone (Volgger, Erschbamer, & Pechlaner, 2021). This opens up the question of which approach for segmenting experiences into episodes works best. Further research into multi-episodic tourism experiences that compares producer-based with visitor-based segmentation is therefore much recommended.

The electrophysiological experience sampling approach to assessing the PE-theory asks for two considerations. First, as mentioned before, one of the drawbacks of measuring physiology in the field is the introduction of noise and error in the form of motion artifacts (Birenboim et al., 2019). This issue was adressed as much as possible by selecting a tourism experience that asks for a minimal amount of movement: sitting down for a 75-minute musical theatre show. In addition, small yet unavoidable motion artifacts have carefully been removed from the data. Still, the theatre show included substantial parts that contained too much motion artifacts due to the nature of the show, which therefore had to be deleted from the data, particularly for 5 scenes. This might affect an accurate representation of these scenes in the eventual analyses, but at the least, it is still able to explain 32% of the variance. This is an increase of roughly 25% as compared to coefficients of determinations that were reported in earlier works on the PE-theory using skin conductance measures (Bastiaansen et al., 2020; Li et al., 2019).

A second consideration of using physiological measures of emotion in studying tourism experiences relates to the still unresolved connection between the physiological and phenomenological dimensions of emotion. Physiological measures of emotion capture bodily processes, which are of a different nature than the contents of our consciousness that are consulted for self-report (Jacobs, 2006). They are hence associated with unique sources of variance, which limits the magnitude of convergence across the two different measures (Mauss & Robinson, 2009). In a review on measures of emotion, Mauss and Robinson (2009) conclude that both physiological and self-reported measures are relevant to understanding emotion as a whole, and should not necessarily be assumed to be interchangeable. Combining physiology and self-report opens up new pathways to a more complete view on how emotions relate to the overall evaluation of experiences in tourism.

A final point of attention relates to the generalization of the results to tourists in general. The sample in this study included both national visitors and international visitors. The characteristics of this sample reflect those of the visitor profile of theme park resorts in general, which in addition to the tourist market for a large part depend on the local resident market as well (Anton Clavé, 2007). While the results from this study can thus be generalized to a mixed audience of both tourists and local residents, generalizations to tourists in general should be made with care. Follow-up research could be aimed at studying differences in the mechanisms of the PE-theory between groups with different background characteristics. As argued by Li (2020), however, the framework of the PE-theory is mostly used to theorize, rather than to generalize. As such, generalization issues do not form a significant issue for the theoretical conclusions as made in this paper.

6. Practical implications

The findings reported in this study confirm that emotionally engaging tourism and leisure participants is a key factor in determining overall experience evaluations. Emotions can be evoked by various factors that are part of the experience designers' toolbox, such as providing interaction, multi-sensory settings and objects, a clear structure or program for the experience, accommodating for the relationships of visitors (i.e. keeping the party composition in mind), and sustaining the action throughout the experience following an animation program (Rossman & Schlatter, 2015). Ma, Scott, Gao, and Ding (2017) found that visitors who attach importance to, are interested in, and pay attention to an experience feel more positive emotions than visitors who show little interest or involvement. Besides allocating resources to create emotionally enhancing experience designs, tourism and leisure managers should thus also devote more resources to marketing and public relation efforts that positively influence such factors as the attached importance and interest in an experience. Also, as suggested by Ma, Gao, Scott, and Ding (2013), tourism and leisure managers could offer opportunities for visitors to celebrate special events or occasions in their facilities, as those visitors are generally more interested in having a salient and memorable experience to begin with.

Note, though, that the findings in this study also imply that visitors should and will not constantly be emotionally engaged throughout the experience. Emotional arousal was significantly related to the overall experience evaluation indeed, but only for 8 of the 17 scenes. Furthermore, some of these 8 scenes were negatively related to the overall experience evaluation. This indicates that for some scenes, an increase in emotional engagement can even negatively affect overall experience evaluations. In line with suggestions from Bastiaansen et al. (2019), tourism and leisure providers should thus carefully determine which emotions should be felt when and subsequently evaluate which segments of the experience are most strongly related to overall experience outcomes and why. Findings can then be used to redesign the respective experience segments in order to enhance the emotional responses in case of a positive correlation with overall experience evaluations, and to mitigate the emotional responses in case of a negative correlation. Arguably, for this evidence-based procedure of experience management and design, methods that allow for capturing the temporal dynamics of the experience (e.g. experience reconstruction or electrophysiological measures such as skin conductance) are indispensable.

7. Conclusion

In sum, the findings in the present study support the notion that emotions form a core factor influencing evaluations of tourism experiences. More particularly, it is shown that overall experience evaluations mostly depend on the pattern of these emotions over time. Using both reconstructed and physiological measures of emotion, it is demonstrated that overall evaluations of a tourism experience are better predicted from the peak and end emotions that are felt during individual episodes of the experience, than from peaks and ends of the experience as a whole. Therefore, further development of the PE-theory should more carefully consider the multi-episodic nature of real-life tourism experiences. Physiological measures of emotion seem to be particularly suited to study these temporal dynamics of tourism experiences.

Funding information

This work was supported by the Centre of Expertise Leisure, Tourism & Hospitality (CELTH) [Storysperience Grant].

References

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

Ariely, D. (1998). Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations. *Journal of Behavioral Decision Making, 11*(1), 19–45. https://doi.org/10.1002/(SICI)1099-0771(199803)11:1<19::AID-BDM277>3.0.CO;2-B

Ariely, D., & Carmon, Z. (2000). Gestalt characteristics of experiences: The defining features of summarized events. *Journal of Behavioral Decision Making, 13*(2), 191–201. https://doi.org/10.1002/(sici)1099-0771(200004/06)13:2<191::Aid-bdm330>3.0.Co;2-a

Ariely, D., & Zauberman, G. (2000). On the making of an experience: The effects of breaking and combining experiences on their overall evaluation. *Journal of Behavioral Decision Making, 13*(2), 219–232.

Bastiaansen, M., Lub, X. D., Mitas, O., Jung, T. H., Passos Ascenção, M., Han, D.-I., … 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

Baumgartner, H., Sujan, M., & Padgett, D. (1997). Patterns of affective reactions to advertisements: The integration of moment-to-moment responses into overall judgments. *Journal of Marketing Research, 34*(2), 219–232. https://doi.org/10.1177/002224379703400203

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

Bigné, J. E., Andreu, L., & Gnoth, J. (2005). The theme park experience: An analysis of pleasure, arousal and satisfaction. *Tourism Management, 26*(6), 833–844. https://doi.org/10.1016/j.tourman.2004.05.006

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
Bosque, I. R., del, & 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
Boswijk, A., Peelen, E., & Olthof, S. (2012). *Economy of experiences*. Amsterdam: European Centre for the Experience and Transformation Economy.
Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). New York, NY: 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
Cetin, G., & Bilgihan, A. (2016). Components of cultural tourists' experiences in destinations. *Current Issues in Tourism, 19*(2), 137–154. https://doi.org/10.1080/13683500.2014.994595
Chajut, E., Caspi, A., Chen, R., Hod, M., & Ariely, D. (2013). In pain thou shalt bring forth children: The peak-and-end rule in recall of labor pain. *Psychological Science, 25*(12), 2266–2271. https://doi.org/10.1177/0956797614551004
Chark, R., King, B., & Tang, C. M. F. (2020). The journey from episode to evaluation: How travelers arrive at summary evaluations. *Journal of Travel Research*, 1–14. https://doi.org/10.1177/004728752098
Clawson, M., & Knetsch, J. L. (1966). *Economics of outdoor recreation*. Baltimore, MD: Johns Hopkins Press.
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
Duerden, M. D., Lundberg, N. R., Ward, P., Taniguchi, S. T., Hill, B., Widmer, M. A., et al. (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. (2017). A theory of structured experience. *Annals of Leisure Research, 22*(1), 97–118. https://doi.org/10.1080/11745398.2017.1312468
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
Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45–55. https://doi.org/10.1037/0022-3514.65.1.45
Geng, X., Chen, Z., Lam, W., & Zheng, Q. (2013). Hedonic evaluation over short and long retention intervals: The mechanism of the peak-end rule. *Journal of Behavioral Decision Making, 26*(3), 225–236. https://doi.org/10.1002/bdm.1755
Godovykh, M., & Tasci, A. (2020). Customer experience in tourism: A review of definitions, components, and measurements. *Tourism Management Perspectives, 39*, 1–10. https://doi.org/10.1016/j.tmp.2020.199694
González-Rodríguez, M. R., Domínguez-Quintero, A. M., & Paddison, B. (2019). The direct and indirect influence of experience quality on satisfaction: The importance of emotions. *Current Issues in Tourism*, 1–19. https://doi.org/10.1080/13683500.2019.1668917
Guerrero-Rodríguez, R., Stepchenkova, S., & Kirilenko, A. (2020). Experimental investigation of the impact of a destination promotional video with physiological and self-reported measures. *Tourism Management Perspectives, 33*. https://doi.org/10.1016/j.tmp.2019.100625
Gutwin, C., Rooke, C., Cockburn, A., Mandryk, R. L., & Lafreniere, B. (2016). *Peak-end effects on player experience in casual games* (Vol. 16). San Jose, CA: Paper presented at the 2016 CHI Conference on Human Factors in Computing Systems - CHI.
Hammitt, W. E. (1980). Outdoor recreation: Is it a multi-phase experience? *Journal of Leisure Research, 12*(2), 107–115. https://doi.org/10.1080/00222216.1980.11969430
Hektner, J. M., Schmidt, J. A., & Csikszentmihaly, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks, CA: Sage Publications.
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* (doctoral dissertation). Wageningen: Wageningen University.
Joo, D., Cho, H., Woosnam, K. M., & Suess, C. (2020). Re-theorizing social emotions in tourism: Applying the theory of interaction ritual in tourism research. *Journal of Sustainable Tourism*, 1–17. https://doi.org/10.1080/09669582.2020.1849237
Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x
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*. New York, NY: Oxford University Press.
Kelly, J. R., & Godbey, G. (1992). Themes and definitions. In J. R. Kelly, & G. Godbey (Eds.), *The sociology of leisure* (pp. 11–28). State College, PA: Venture Publishing Inc.
Kemp, S., Burt, C. D. B., & Furneaux, L. (2008). A test of the peak-end rule with extended autobiographical events. *Memory & Cognition, 36*(1), 132–138. https://doi.org/10.3758/mc.36.1.132
Kensinger, E. A. (2009). Remembering the details: Effects of emotion. *Emot Rev, 1*(2), 99–113. https://doi.org/10.1177/1754073908100432
Kensinger, E. A., & Schacter, D. L. (2008). Memory and emotion. In M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 601–617). New York, NY: The Guilford 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
Kim, H., & Kim, B. (2019). The evaluation of visitor experiences using the peak-end rule. *Journal of Heritage Tourism*, 1–13. https://doi.org/10.1080/1743873x.2019.1575388
Kim, S.-B., Kim, D.-Y., & Bolls, P. (2014). Tourist mental-imagery processing: Attention and arousal. *Annals of Tourism Research, 45*, 63–76. https://doi.org/10.1016/j.annals.2013.12.005
Kirchberg, V., & Tröndle, M. (2015). The museum experience: Mapping the experience of fine art. *Curator: The Museum Journal, 58*(2), 169–193. https://doi.org/10.1111/cura.12106
Larsen, R. J., & Fredrickson, B. L. (1999). Measurement issues in emotion research. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 155–184). New York, NY: Russell Sage Foundation.
Lee, Y.-J. (2016). The relationships amongst emotional experience, cognition, and behavioural intention in battlefield tourism. *Asia Pacific Journal of Tourism Research, 21*(6), 697–715. https://doi.org/10.1080/10941665.2015.1068195
Li, S. (2020). Using self-report and skin conductance measures to evaluate theme park experieces. *Journal of Vacation Marketing*, 1–18. https://doi.org/10.1177/13567667720969738
Liersch, M. J., & McKenzie, C. R. M. (2009). Duration neglect by numbers—and its elimination by graphs. *Organizational Behavior and Human Decision Processes, 108*(2), 303–314. https://doi.org/10.1016/j.obhdp.2008.07.001
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
Li, S., Walters, G., Packer, J., & Scott, N. (2018a). A comparative analysis of self-report and psychophysiological measures of emotion in the context of tourism advertising. *Journal of Travel Research, 57*(8). https://doi.org/10.1177/0047287517733555
Li, S., Walters, G., Packer, J., & Scott, N. (2018b). Using skin conductance and facial electromyography to measure emotional responses to tourism advertising. *Current Issues in Tourism, 21*(15), 1761–1783. https://doi.org/10.1080/13683500.2016.1223023
Li, S., Walters, G., Packer, J., & Scott, N. (2019). Using facial electromyography to test the peak-end rule in tourism advertising. *Journal of Hospitality & Tourism Research*, 1–23.
Ma, J., Gao, J., Scott, N., & Ding, P. (2013). Customer delight from theme park experiences: The antecedents of delight based on cognitive appraisal theory. *Annals of Tourism Research, 42*, 359–381. https://doi.org/10.1016/j.annals.2013.02.018
Ma, J., Scott, N., Gao, J., & Ding, P. (2017). Delighted or satisfied? Positive emotional responses derived from theme park experiences. *Journal of Travel & Tourism Marketing, 34*(1), 1–19. https://doi.org/10.1080/10548408.2015.1125824
Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion, 23*(2), 209–237. https://doi.org/10.1080/02699930802204677
Miron-Shatz, T. (2009). Evaluating multiepisode events: Boundary conditions for the peak-end rule. *Emotion, 9*(2), 206–213. https://doi.org/10.1037/a0015295
Mitas, O., Mitasova, H., Millar, G., Boode, W., Neveu, V., Hover, M., , … Bastiaansen, M., et al. (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
Mitas, O., Yarnal, C., Adams, R., & Ram, N. (2012). Taking a "peak" at leisure travelers' positive emotions. *Leisure Sciences, 34*(2), 115–135. https://doi.org/10.1080/01490400.2012.652503
Mommaas, H. (2000). *De vrijetijdsindustrie in stad en land: Een studie naar de markt van belevenissen*. The Hague: Sdu Uitgevers.
Moss, S. (2010). *An introduction to the entertainment industry*. Wallingford: CABI.
Müller, U. W. D., Witteman, C. L. M., Spijker, J., & Alpers, G. W. (2019). All's bad that ends bad: There is a peak-end memory bias in anxiety. *Frontiers in Psychology, 10*, 1272. https://doi.org/10.3389/fpsyg.2019.01272
Nawijn, J. (2010). The holiday happiness curve: A preliminary investigation into mood during a holiday abroad. *International Journal of Tourism Research, 12*(3), 281–290. https://doi.org/10.1002/jtr.756
Nawijn, J., & Fricke, M.-C. (2015). Visitor emotions and behavioral intentions: The case of concentration camp memorial Neuengamme. *International Journal of Tourism Research, 17*(3), 221–228. https://doi.org/10.1002/jtr.1977
Nawijn, J., Marchand, M. A., Veenhoven, R., & Vingerhoets, A. J. (2010). Vacationers happier, but most not happier after a holiday. *Appl Res Qual Life, 5*(1), 35–47. https://doi.org/10.1007/s11482-009-9091-9
Nawijn, J., Mitas, O., Lin, Y., & Kerstetter, D. (2012). How do we feel on vacation? A closer look at how emotions change over the course of a trip. *Journal of Travel Research, 52*(2), 265–274. https://doi.org/10.1177/0047287512465961
Newman, D. B., Schwarz, N., & Stone, A. A. (2020). Global reports of well-being overestimate aggregated daily states of well-being. *The Journal of Positive Psychology*, 1–10. https://doi.org/10.1080/17439760.2020.1725608
Osborne, T., & Jones, P. I. (2017). Biosensing and geography: A mixed methods approach. *Applied Geography, 87*, 160–169. https://doi.org/10.1016/j.apgeog.2017.08.006
Park, S., Hahn, S., Lee, T., & Jun, M. (2018). Two factor model of consumer satisfaction: International tourism research. *Tourism Management, 67*, 82–88. https://doi.org/10.1016/j.tourman.2018.01.007
Paül i Agustí, D., Rutllant, J., & Lasala Fortea, J. (2019). Differences in the perception of urban space via mental maps and heart rate variation (HRV). *Applied Geography, 112*. https://doi.org/10.1016/j.apgeog.2019.102084

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Redelmeier, D. A., Katz, J., & Kahneman, D. (2003). Memories of colonoscopy: A randomized trial. *Pain, 104*(1), 187–194. https://doi.org/10.1016/s0304-3959(03)00003-4

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–55.

Rossman, J. R., & Schlatter, B. E. (2015). *Recreation programming: Designing leisure experiences*. Champaign, IL: Sagamore.

Rozin, A., Rozin, P., & Goldberg, E. (2004). The feeling of music past: How listeners remember musical affect. *Music Perception, 22*(1), 15–39.

Schäfer, T., Zimmermann, D., & Sedlmeier, P. (2014). How we remember the emotional intensity of past musical experiences. *Frontiers in Psychology, 5*, 911. https://doi.org/10.3389/fpsyg.2014.00911

Schneider, S., Stone, A. A., Schwartz, J. E., & Broderick, J. E. (2011). Peak and end effects in patients' daily recall of pain and fatigue: A within-subjects analysis. *The Journal of Pain, 12*(2), 228–235. https://doi.org/10.1016/j.jpain.2010.07.001

Schreiber, C. A., & Kahneman, D. (2000). Determinants of the remembered utility of aversive sounds. *Journal of Experimental Psychology, 129*(1), 27–42.

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

Scott, N., & Le, D. (2017). Tourism experience: A review. In N. Scott, J. Gao, & J. Ma (Eds.), *Visitor experience design* (pp. 30–49). Wallingford: CABI.

Seta, J. J., Haire, A., & Seta, C. E. (2008). Averaging and summation: Positivity and choice as a function of the number and affective intensity of life events. *Journal of Experimental Social Psychology, 44*(2), 173–186. https://doi.org/10.1016/j.jesp.2007.03.003

Shoval, N., Schvimer, Y., & Tamir, M. (2018a). Real-time measurement of tourists' objective and subjective emotions in time and space. *Journal of Travel Research, 57*(1), 3–16. https://doi.org/10.1177/0047287517691155

Shoval, N., Schvimer, Y., & Tamir, M. (2018b). Tracking technologies and urban analysis: Adding the emotional dimension. *Cities, 72*, 34–42. https://doi.org/10.1016/j.cities.2017.08.005

Stewart, W. P. (1998). Leisure as multiphase experiences: Challenging traditions. *Journal of Leisure Research, 30*(4), 391–400. https://doi.org/10.1080/00222216.1998.11949840

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*. https://doi.org/10.1016/j.jdmm.2019.100408

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

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). *Automatic identification of artifacts in electrodermal activity data*. Paper presented at the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Milan, Italy: EMBC).

Theil, H. (1961). *Economic forecasts and policy*. Amsterdam: North-Holland Publishing Company.

Tröndle, M., Greenwood, S., Kirchberg, V., & Tschacher, W. (2012). An integrative and comprehensive methodology for studying aesthetic experience in the field. *Environment and Behavior, 46*(1), 102–135. https://doi.org/10.1177/0013916512453839

Volgger, M., Erschbamer, M., & Pechlaner, H. (2021). Destination design: New perspectives for tourism destination development. *Journal of Destination Marketing & Management*, 1–6. https://doi.org/10.1016/j.jdmm.2021.100561

Volo, S. (2021). The experience of emotion: Directions for tourism design. *Annals of Tourism Research, 86*, 1–11. https://doi.org/10.1016/j.annals.2020.103097

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