Universität
Basel

Fakultät für
Psychologie

Measuring and Understanding User Experience: Current State and Future Needs

Inauguraldissertation zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Sebastian A. C. Perrig

aus Basel (BS), Schweiz

Basel, 2023

Universität
Basel

Fakultät für
Psychologie

FAKULTÄT FÜR PSYCHOLOGIE 2003 UNIVERSITÄT BASEL

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Klaus Opwis (Erstgutachter)

Dr. Javier Andrés Bargas-Avila (Zweitgutachter)

Datum des Doktoratsexamen:

Dekan:in der Fakultät für Psychologie

Universität
Basel

Fakultät für
Psychologie

FAKULTÄT FÜR PSYCHOLOGIE 2003 UNIVERSITÄT BASEL

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt. Es handelt sich dabei um folgende Manuskripte:

1. Perrig, S. A. C., Ueffing, D., Opwis, K., & Brühlmann, F. (2023). Smartphone app aesthetics influence users' experience and performance. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1113842
2. Perrig, S. A. C., von Felten, N., Honda, M., Opwis, K., & Brühlmann, F. (2023). Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-Pos. *International Journal of Human–Computer Interaction*. https://doi.org/10.1080/10447318.2023.2258634
3. Aeschbach, L. F., Perrig, S. A. C., Weder, L., Opwis, K., & Brühlmann, F. (2021). Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY). https://doi.org/10.1145/3474660
4. Perrig, S. A. C., Scharowski, N., Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2023). Independent validation of the Player Experience Inventory: Findings from a large set of video game players. *Manuscript submitted for publication*.

Spezifizierung des eigenen Forschungsbeitrags zu den Manuskripten:

1. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization
☒ Data curation
☒ Formal Analysis
☐ Funding acquisition
☒ Investigation
☒ Methodology
☐ Project administration
☐ Resources
☐ Software
☐ Supervision
☐ Validation
☒ Visualization
☒ Writing – original draft
☒ Writing – review & editing

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

2. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization
☒ Data curation
☒ Formal Analysis
☐ Funding acquisition
☒ Investigation
☒ Methodology
☒ Project administration
☐ Resources
☐ Software
☐ Supervision
☐ Validation
☒ Visualization
☒ Writing – original draft
☒ Writing – review & editing

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

3. Eigener Beitrag nach CRediT[1]:

☒ Conceptualization
☐ Data curation
☐ Formal Analysis
☐ Funding acquisition
☒ Investigation
☒ Methodology
☐ Project administration
☐ Resources
☐ Software
☐ Supervision
☐ Validation
☐ Visualization
☐ Writing – original draft
☒ Writing – review & editing

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

[1] https://casrai.org/credit/

4. Eigener Beitrag nach CRediT[1]:

- ☒ Conceptualization
- ☒ Data curation
- ☒ Formal Analysis
- □ Funding acquisition
- ☒ Investigation
- ☒ Methodology
- ☒ Project administration
- □ Resources
- □ Software
- □ Supervision
- □ Validation
- ☒ Visualization
- ☒ Writing – original draft
- ☒ Writing – review & editing

Das Manuskript wurde bisher für keine anderen Qualifikationsarbeiten eingereicht.

Universität
Basel

Fakultät für
Psychologie

Open-Science Aspekte der Manuskripte:

1. Preregistration: □ ja ☒ nein

   Open-Access-Publikation: ☒ ja □ nein

   Open-Access-Data/Analyse: ☒ ja □ nein

   Ort/URL der Daten und Analysen: https://osf.io/qevpk/

2. Preregistration: ☒ ja □ nein

   Open-Access-Publikation: ☒ ja □ nein

   Open-Access-Data/Analyse: ☒ ja □ nein

   Ort/URL der Daten und Analysen: https://osf.io/efy4s/

3. Preregistration: □ ja ☒ nein

   Open-Access-Publikation: ☒ ja □ nein

   Open-Access-Data/Analyse: ☒ ja □ nein

   Ort/URL der Daten und Analysen: https://osf.io/4xz2v/

4. Preregistration: ☒ ja □ nein

   Open-Access-Publikation: ☒ ja □ nein

   Open-Access-Data/Analyse: ☒ ja □ nein

   Ort/URL der Daten und Analysen: https://osf.io/8xuhr/

Ort, Datum .... Basel, 11. April 2024

Signatur .... S.Perrig

Vorname Nachname .... Sebastian A. C. Perrig

Contents

Abstract **8**

Introduction **9**

Theoretical background **12**

What is UX? 12

Using scales to measure UX 13

Investigating the quality of survey scales 14

Theory as a foundation for measurement 16

Aesthetics in HCI 18

Research transparency 20

Player experience 22

Summary of the manuscripts **24**

First manuscript: "Smartphone app aesthetics influence users' experience and performance" 27

Second manuscript: "Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-pos" 31

Third manuscript: "Transparency in measurement reporting: A systematic literature review of CHI PLAY" 36

Fourth manuscript: "Independent validation of the Player Experience Inventory: Findings from a large set of video game players" 40

General discussion **44**

The importance of theory as a foundation for measurement 46

Scale validation as an ongoing process 47

Measuring and understanding UX 48

Limitations and future directions 50

Conclusion **51**

Acknowledgments **67**

Appendix **70**

Abstract

Over the last two decades of Human-Computer Interaction, the research focus has shifted from usability to the more holistic concept of user experience (UX). Given the subjective nature of UX, emphasis is placed on psychological constructs which are not directly observable. Thus, methods are needed to adequately measure these complex experiences, among which self-reported survey scales are some of the most popular. However, the psychometric quality of those scales is not always known or (re-)investigated, which would be essential to have confidence in the findings obtained from them. Consequently, this thesis consists of four manuscripts investigating the measurement of UX from different perspectives and through the investigation of various survey scales. In particular, the first manuscript studies one crucial question from UX research, namely, how interface aesthetics affect users' subjective experience and their objective performance in the to-date understudied but increasingly popular context of smartphone devices. The second manuscript looks at the quality of a particular scale for measuring website aesthetics and considers the effect negatively formulated items have on scale quality by offering an alternative positive version. The third manuscript investigates how survey scales are currently used and reported on in a particular area of research, namely player experience (PX). Finally, the fourth manuscript examines the quality of one particular scale for measuring PX. Overall, this thesis provides insights into the current state of scale usage as a method in UX research while highlighting areas for future improvement.

Introduction

User experience (UX) has become a central concept in Human-Computer Interaction (HCI) over the last two decades, with the field moving from a focus on usability to UX as a more holistic view of interaction (Hassenzahl, 2018; Pettersson et al., 2018). Modern UX research thus has broadened its focus from task completion and pragmatic goals to consider additional aspects critical to a positive experience with technology, such as aesthetics or the satisfaction of user needs (Hassenzahl & Tractinsky, 2006). These experiences can be complex and, more importantly, are subjective in their nature. Hence, methods are needed to measure the subjective constructs essential to UX. In empirical psychology, the term "construct" is typically used to refer to aspects such as psychological traits or abilities of individuals that are not directly observable (Hopkins, 1998). More than just observing people's behavior is needed to capture relevant constructs fully, and other methods are required to measure the subjective experience indirectly (DeVellis, 2017). Among those methods used to study subjective experiences, survey scales, sometimes called questionnaires, are one of the most popular in UX research (Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018).

At their core, survey scales use a set of statements (i.e., items) to indirectly gauge the extent of a construct (DeVellis, 2017). Within the social sciences, there is a long tradition of using survey scales for measurement (Gault, 1907). However, recent research in UX-related areas, such as social and personality psychology, has shown that the psychometric quality of survey scales is not always given when re-investigated, even if these scales were previously validated (Hussey & Hughes, 2020), and ongoing efforts of validation are rarely reported (Flake et al., 2017). Concerning HCI, independent validation of previously published survey scales has also raised doubts about their proposed psychometric quality and theoretical models (e.g., Kayser et al., 2021; Memeti et al., 2022; Perrig, Scharowski, & Brühlmann, 2023; von Felten et al., 2022). Further, past work has demonstrated that UX researchers frequently resort to ad hoc scales (i.e., self-developed scales typically created for use in a single study), which frequently are of questionable psychometric quality, if their quality is even investigated (Bargas-Avila &

Hornbæk, 2011). These findings raise the question of whether survey scales currently used to measure UX are adequate and where there is room for improvement.

In addition, past research has questioned whether survey scales are appropriate to measure UX in the first place. In particular, some UX researchers have exhibited skepticism towards the process of breaking down UX into constructs to measure it, considering it to be an oversimplification, contradicting the holistic nature of UX (Law et al., 2014). In contrast, Cairns and Power (2018) suggested that this contradiction between a holistic view of UX and the reductionist measurement process can be acknowledged through the proper use of psychometric methods. For them, modern statistical methods allow for richer interpretations of the data gathered using survey scales, thus providing a differentiated picture of UX. However, the crucial point here is the "proper use" of psychometrics. As mentioned above, survey scales used in research frequently are of questionable quality, thus requiring further attention. In addition, survey scales are not only one of the most popular but also among the most frequently misused research methods, both in HCI and other fields of research (Green et al., 2008; Vermeeren et al., 2010). One reason is that researchers face numerous decisions during the measurement process, allowing for flexibility, which can lead to questionable measurement practices (Flake & Fried, 2020). Without proper guidance, researchers thus risk compromising the validity of their collected results and, consequently, the findings derived from them. Therefore, it is essential to know if UX researchers currently follow appropriate practices and, if not, how they can be supported in doing so in the future.

This thesis deals with the above-mentioned issues in different ways, either directly or indirectly, and through selected survey scales and areas of UX, illustrating the current state of measurement using survey scales in UX research. The first manuscript focused on a vital construct for UX, aesthetics, and investigated how interface aesthetics influence the subjective experience of users and their objective performance in the increasingly popular (Tenzer, 2023) but understudied context of smartphone devices. The second manuscript re-investigated the quality of and theoretical model behind a previously published scale for measuring aesthetics, the Visual Aesthetics of Websites Inventory (VisAWI, Moshagen & Thielsch, 2010). Furthermore, it studied how reverse-coded items commonly recommended to be used (Nunnally, 1978) can affect the quality of survey scales, distorting the results collected with them. The third manuscript examined how survey scales are currently used and reported on in one particular area of UX research: player experience (PX). Based on these findings, we formulated recommendations for more transparent measurement reporting. Finally, the fourth manuscript investigated the quality of one recently proposed survey scale for measuring constructs relevant to PX, the Player Experience Inventory (PXI, Abeele et al., 2020). Jointly, these manuscripts show how survey scales are currently used in UX research, how they should be used, and where there is room for improvement for future research working with this vital method.

Theoretical background

What is UX?

There is a plethora of definitions attempting to establish what UX is. For example, the website of the Experience Research Society[2] contains a non-exhaustive collection of 27 proposed definitions of UX collected from various academic and non-academic sources. Among these, one of the more detailed definitions by Hassenzahl and Tractinsky (2006) characterizes UX as "a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organizational/social setting, meaningfulness of the activity, voluntariness of use, etc.)" (p. 6). In another definition, the International Organization for Standardization (ISO) describes UX as a "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" (International Organization for Standardization, 2019).

While both of these definitions share the implicit idea that UX focuses on subjective aspects of the user's interaction with a system, something that most UX researchers and practitioners agree on (Law et al., 2009), they also show that UX is a broad concept, encompassing many aspects relevant to the interaction. After over two decades of UX research, there appears to be limited consensus on defining and conceptualizing UX. In this regard, previous work has noted a lack of conceptual clarity in the transition from usability to UX (Miki, 2013). This is likely influenced by the multidisciplinary nature of UX research, resulting in numerous perspectives and definitions of UX that have turned it into a broad concept, consequently limiting its use for practice and research and making it hard to understand and too expansive in terms of its scope (Roto, 2009). This broadness is also reflected in the fact that UX can be considered an umbrella construct (Tractinsky, 2018). According to Hirsch and Levin (1999), umbrella constructs represent "a broad concept or idea used loosely to encompass and account for

[2] https://experienceresearchsociety.org/ux/ux-definitions/, accessed on December 8, 2023.

a set of diverse phenomena" (p. 200). The main issue with these umbrella constructs is that researchers find it difficult or even impossible to agree on how they should be defined and measured (Tractinsky, 2018). Thus, UX as an umbrella constructs can be hard to grasp, making it difficult to understand what is (and is not) part of UX. Moreover, it is easier for research to build upon common ground and previous findings with a clear understanding of what UX is and how its components should be measured. The fragmented nature of HCI research and a lack of theory building has been a point of critique in the past (Kostakos, 2015; Oulasvirta & Hornbæk, 2016). Thus, there is a need for research to understand better what aspects of UX, or rather which constructs, are essential and how these constructs relate to one another. Furthermore, more research is needed looking at how constructs of UX can be operationalized and how these insights can lead to more comprehensive frameworks, models, and theories of UX. In this context, survey scales can play an essential role in measuring UX.

Using scales to measure UX

Given the subjective nature of UX, researchers have to resort to various methods to study it, among which survey scales are one of the most popular (Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018). According to DeVellis (2017), survey scales are "collections of items combined into a composite score and intended to reveal levels of theoretical variables not readily observable by direct means" (p. 15). Thus, survey scales offer researchers a way to measure constructs that are not directly observable. For this, participants respond to a series of statements (i.e., items) designed to capture various aspects of the target construct because it is assumed that the construct's level or magnitude impacts the responses to these statements (DeVellis, 2017). Answers are collected using pre-defined response options like a Likert-type agreement scale (e.g., 1 - 7) or a semantic differential format (e.g., "good" - "bad"). After data collection, participant responses to these items are typically computed into scores (e.g., averaging them) to create an estimate (i.e., the score) representing the not directly observable target construct. However, these scores can only be trusted as representative proxies of the target construct if the scales are correctly used, are of adequate psychometric quality, and have been developed with sufficient care (Flake & Fried, 2020). However, past reviews of UX research have shown that a large proportion of employed scales are ad hoc scales (Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018). The issue with utilizing such ad hoc or self-developed scales is that they, most times, are of unknown or questionable psychometric quality. Developing a high-quality survey scale requires considerable work. It includes procedures to ensure that items are carefully formulated based on theoretical considerations and efforts to investigate the psychometric quality of the developed scale (DeVellis, 2017). As part of an iterative development process, a scale and its items are re-investigated and revised until the final version of the scale is formed (Furr, 2011). Researchers can then use the scale across studies, ideally without modifying it (e.g., not changing item wording or the response format) (Juniper, 2009). This development process is usually given less attention for ad hoc scales, typically designed for one particular study and only used once. In such cases, the scales and their corresponding items are often assembled with limited theoretical considerations guiding the design process (DeVellis, 2017).

Investigating the quality of survey scales

As mentioned, scale validation is an ongoing and iterative process. The psychometric quality of a scale should ideally be re-investigated whenever it is used, or at least when used with samples drawn from new participant populations (Furr, 2011). When investigating the quality of scales, researchers are typically concerned with three quality criteria: *objectivity, reliability*, and *validity* (DeVellis, 2017).

Objectivity. Objectivity means that "any statement of fact made by one scientist should be independently verifiable by other scientists" (Nunnally, 1978, p. 6). Moosbrugger and Kelava (2000) distinguish three forms of objectivity: objectivity of implementation, objectivity of evaluation, and objectivity of interpretation. While objectivity of implementation can be increased through standardization (e.g., clear instructions on how the scale should be filled out), objectivity of evaluation is improved if there are clear guidelines on how to assess the participants’ responses (e.g., clearly describing how to code answers). Finally, objectivity of interpretation is raised by providing norm data to compare the results of participants to, for example, based on past ratings by other participants from the same target population (Moosbrugger & Kelava, 2000). In summary, objectivity can be enhanced by supplying clear instructions and guidelines on how to use the scale and interpret the results and by using survey scales consistently across research without deviation from the original wording of the items and instructions while following the guidelines provided (e.g., from a manual accompanying the scale).

Reliability. The second quality criterion, reliability, asks "how accurately a test measures the thing which it does measure" (Kelley, 1927, p. 14). Here, the goal is to ensure that scores gathered by using a survey scale are not tainted by measurement error but are as consistent as possible in measuring the target construct (Moosbrugger & Kelava, 2000). Two indicators of internal consistency, coefficients alpha (Cronbach, 1951) and omega (McDonald, 1999), are commonly recommended (Dunn et al., 2014), although further methods exist (e.g., test-retest or split-half reliability). Given that both objectivity and reliability aim to increase the consistency of scale usage across research, and consequently, the findings derived from data gathered using those survey scales, they are crucial to the ongoing discussions in psychology regarding a "replication crisis" (e.g., Open Science Collaboration, 2015). Furthermore, because measurement errors are undesirable due to their negative impact on scale quality, efforts to improve data quality (e.g., Brühlmann et al., 2024; Brühlmann et al., 2020) are essential for research using survey scales.

Validity. Finally, validity considers "whether a test really measures what it purports to measure" (Kelley, 1927, p. 14). While reliability ensures that variations in scale ratings are due to the true value of some construct rather than due to error, validity assesses whether the measured construct actually is the construct of interest (DeVellis, 2017). Thus, validity is the most crucial quality criterion for a scale, although objectivity and reliability are necessary precursors to it (Moosbrugger & Kelava, 2000). According to DeVellis (2017), validity can be differentiated into content validity, criterion-related validity, and construct validity: Content validity requires that a scale's items ought to reflect the conceptual definition appropriate to the scale's content domain (e.g., by selecting items for a trust scale that cover all relevant aspects of trust brought up in a definition). Thus, a scale's theoretical foundation is essential here, as this quality criterion is closely linked to the definition of the target construct(s). In contrast, criterion-related validity (or predictive validity) assesses whether the empirical relationship between a scale or item and an external criterion is as expected (e.g., by assessing if ratings from a subjective trust scale can predict if users follow a system's recommendations). Therefore, criterion-related validity is not directly concerned with the theoretical underpinnings of a scale. Lastly, construct validity places the scale in relation to other measures (e.g., other scales measuring related constructs) to investigate if the relationships between the scale and these other measures are as would be expected based on theory, given that the scale actually measures the intended construct (e.g., ratings from a trust scale should correlate highly with other measures of trust, but to a lesser extend with other UX constructs such aesthetics or usability). Hence, this criterion is directly concerned with the theoretical relationships between the scale's construct(s) and other measures (DeVellis, 2017). Numerous statistical methods exist to provide evidence for a scale's validity (for an overview, see chapter 2.4.2 in Moosbrugger & Kelava, 2000), including exploratory and confirmatory factor analysis (EFA and CFA) or methods investigating the relationship of the measured construct to other constructs or associated variables (e.g., through correlations). However, given that validity considers what the scale aims to measure (i.e., the actual target construct), theoretical reasoning is essential to the validation process and for interpreting results from these statistical analyses.

Theory as a foundation for measurement

Scales are typically developed based on theoretical models, representing the researchers' understanding of the target construct. Items of a scale are formulated following theoretical considerations to represent as many aspects of the target construct as possible (DeVellis, 2017), and in the case of multi-dimensional scales, theoretical reasoning is needed to formulate how many and which constructs are measured by a scale and how these constructs are related to one another (Brown, 2015). Given these theoretical considerations, researchers either explicitly or implicitly commit to the underlying theory behind a scale when employing it (DeVellis, 2017). Consequently, a lack of standardized and agreed-upon measures for a construct results in numerous (conscious or unconscious) theoretical assumptions, limiting common ground on which aspects of the construct to measure. Furthermore, scales with varying theoretical conceptualizations of the same target construct can complicate the comparability of research findings. For example, let us consider a case where one researcher measures trust using a scale that assumes trust consists of two constructs (trust and distrust). In contrast, another researcher uses a trust scale that assesses three constructs (capability, benevolence, and integrity). In this case, both researchers will believe that they are measuring trust. However, when comparing their results, they will likely differ, not necessarily because of actual effects, but rather due to the difference in theory and the resulting constructs measured. Concerning UX research, the argument has been made that inconsistencies among findings on the same constructs (i.e., usability and aesthetics) are due to methodological inconsistencies rather than because of actual effects (Hassenzahl & Monk, 2010). Thus, a theoretical basis is quintessential when working with survey scales.

In this regard, Maul (2017) has highlighted the importance of sound theoretical foundations for survey scales. Across three studies, they demonstrated that survey scales consisting of items with nonsensical but consistent wording (e.g., negative vs. positive items) can still deliver satisfactory results in commonly used methods to assess reliability and validity. Therefore, researchers can not rely on results from statistical analyses alone when developing and investigating the quality of survey scales because the statistical methods themselves do not account for the theory behind the scales but rather just detects patterns in the data. Instead, researchers also have to account for theoretical considerations when choosing their scales and interpreting their results, for example, by paying attention to the definitions of the measured constructs (Maul, 2017).

The work by Maul (2017) emphasizes how the quality of a survey scale goes beyond the mere psychometric analysis of the three criteria: objectivity, reliability, and validity. In addition to the methods for statistically analyzing the quality of a scale, such as factor analysis or indicators of internal consistency, researchers have to critically engage with the measures they are employing and the theoretical foundations of those measures. Given the aforementioned concerns regarding a lack of common ground and theory in HCI (Kostakos, 2015; Oulasvirta & Hornbæk, 2016), this further highlights the importance of critical research into survey scales and their theoretical underpinnings.

In summary, survey scales are among the most common methods used to measure UX and its constructs. Adequately developed survey scales have many advantages, such as allowing researchers to measure constructs that are not directly observable and the possibility of scientific comparison and generalization. Some, if not all, of these advantages are lost when scales of unknown or improper quality are commonly used or when the same high-quality scales are not used across research. Consequently, numerous researchers have advocated the use of standardized scales in HCI research rather than self-developed measures (e.g., Hornbæk, 2006; Hornbæk & Law, 2007; Sauro & Lewis, 2009). Moreover, scales ought to be developed based on theoretical models, and the adequacy of the scales and their underlying models needs to be re-investigated in an ongoing process of psychometric quality investigation.

Aesthetics in HCI

Within UX research, aesthetics is one of the most frequently studied constructs (Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018). In this context, aesthetics is often treated as a synonym of beauty and defined "as an immediate pleasurable subjective experience that is directed toward an object and not mediated by intervening reasoning" (Moshagen & Thielsch, 2010, p. 690).

Starting in the 1990s, HCI researchers began to investigate how the aesthetic appeal of a product affects users' perceptions of the interface's functional aspects and thus its usability (Kurosu & Kashimura, 1995; Tractinsky, 1997). This shift beyond usability meant that much early work on UX focused on how aesthetics and usability interact. In this regard, a seminal study by Tractinsky et al. (2000) suggested that "what is beautiful is usable," postulating a halo effect of aesthetics: an inference from the aesthetic design of the system to other attributes, in this case, the system's usability. In the following years, numerous researchers revisited the relationship between usability and aesthetics, including Tuch et al. (2012), who found that the halo effect of aesthetics is reversed under certain conditions (i.e., "what is usable is beautiful"). Overall, however, findings on the connection between usability and aesthetics were somewhat inconclusive, with correlations between the two constructs reported across studies ranging from non-existing to almost perfect (Hassenzahl & Monk, 2010), and research is still ongoing (e.g., Minge & Thüring, 2018; Schrepp et al., 2021). To possibly explain these inconsistent findings, Hassenzahl and Monk (2010) suggested that they were due to methodological differences among studies, such as discrepancies in how aesthetics and usability are measured or which experimental stimuli are employed in a study.

Beyond investigating the relationship between aesthetics and usability, there are also studies in HCI research on how aesthetics can influence other aspects of UX such as a system's overall appeal (e.g., Hausman & Siekpe, 2009), user satisfaction (e.g., Seng & Mahmoud, 2020), user preference (e.g., Lee & Koubek, 2010), intention to use (e.g., Pengnate et al., 2019), system trustworthiness (e.g., Skulmowski et al., 2016), and user emotion (e.g., Bhandari et al., 2019). In addition, studies have investigated how aesthetics affect users' objective performance when interacting with a system. In this regard, Thielsch et al. (2019) conducted a meta-analysis, demonstrating a small positive effect of interface aesthetics on performance ($g = 0.12$). However, the authors also stressed a need for more high-quality research addressing the effect of aesthetics on performance. Furthermore, research on the issue thus far did not include smartphone devices, despite their increasing popularity (Tenzer, 2023), something that was addressed in the first manuscript of this thesis (Perrig, Ueffing, et al., 2023).

Despite aesthetics' importance to HCI, researchers frequently resort to ad hoc or single-item scales to measure it (Abbas et al., 2022; Thielsch et al., 2019). Moreover, few validated survey scales of aesthetics exist (Moshagen & Thielsch, 2010), reflecting the general challenges of scale use as outlined above. Two exceptions for validated questionnaires of aesthetics are the scale for the measurement of classic and expressive aesthetics by Lavie and Tractinsky (2004) and the VisAWI by Moshagen and Thielsch (2010), with the latter building upon the work of the former. However, the VisAWI was initially developed in German and was never validated in English. Furthermore, findings from the first manuscript of this thesis suggested that the inclusion of reverse-coded or negatively formulated items undermines the psychometric quality of the VisAWI, which has been reported for other scales in HCI as well (e.g., Lewis et al., 2013; Perrig, Scharowski, & Brühlmann, 2023; Sauro & Lewis, 2011). Together, these two reasons motivated the second manuscript (Perrig, von Felten, et al., 2023), reporting on developing an alternative positive-item-only version of the VisAWI and validating the English version of the scale.

In summary, past HCI research has shown aesthetics to be an essential factor for UX, interacting with and affecting numerous other parts of user perception and interaction. Despite this importance, measuring aesthetics has been and continues to be an issue in HCI research, possibly leading to inconclusive findings requiring further attention.

Research transparency

When conducting a study, researchers face numerous decisions, with each choice possibly affecting the final results and the conclusions derived from them. This "garden of forking paths" (Gelman & Loken, 2014) comes with the risk of researchers going down problematic paths and intentionally or unintentionally engaging in questionable research practices (QRPs). These QRPs, such as hypothesizing after results are known (HARKing) or the manipulation of experimental and analytical methods in pursuit of statistically significant results (p-hacking) (Cockburn et al., 2020), undermine the conclusions of a study (Flake & Fried, 2020). QRPs have received much attention in recent research in HCI and related fields. In psychology, the concern of a replication crisis has been raised, demonstrating that many published research findings are not reproducible, potentially due to QRPs (Open Science Collaboration, 2015). Concerning HCI, an article by Cockburn et al. (2020) has discussed the possibility of such QRPs within computer science, arguing that many areas of computer science, including HCI, are also susceptible to a replication crisis. Remedies against QRPs frequently aim at increasing research transparency, for example, by pre-registering studies (Cockburn et al., 2018), by encouraging researchers to share their materials (e.g., data and analysis scripts) (Wacharamanotham et al., 2020), or by having journal guidelines that demand increased transparency and openness (Ballou et al., 2021).

Concerning measurement, researchers face comparable amounts of flexibility, which can lead to questionable measurement practices (QMPs). According to Flake and Fried (2020), QMPs are "decisions researchers make that raise doubts about the validity of measure use in a study, and ultimately the study's final conclusions" (p. 2). While no definitive list of QMPs exists, they cover a range of issues, from researcher ignorance or negligence to intentionally misleading practices, such as not defining the constructs being measured or adapting the items of a standardized survey scale without clear reasoning to do so or without clearly reporting what was adapted (Flake & Fried, 2020). Similar to QRPs, QMPs are usually associated with nontransparent reporting, both intentional or unintentional, which obscures researcher decisions regarded as questionable, likewise threatening the validity of research results and the conclusions derived from them (Flake & Fried, 2020). Furthermore, researchers providing only limited information on their measurement process makes it challenging to determine which QMPs exist and how frequent QMPs are within a particular field of research. While past work has shown indications for the existence of QMPs in UX research, such as frequent use of self-developed scales (Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018) and nontransparent reporting of the scale items used (Bargas-Avila & Hornbæk, 2011), little is currently known about the state of transparency of measurement reporting in HCI. This motivated the third manuscript of this thesis (Aeschbach et al., 2021), which investigated how transparently PX researchers report on their measurement process.

Player experience

Given the popularity of digital games, the study of PX, "user experience in the specific context of digital games" (Nacke & Drachen, 2011, p. 1), has emerged as a thriving area of research within HCI. In this regard, the concept of UX is applied under the term PX to cover the characteristics specific to the interaction with digital, and at times non-digital (e.g., Rogerson et al., 2018), games. Accordingly, PX "denotes the individual and personal experience of playing games" (Wiemeyer et al., 2016, p. 246). Given this focus on players' interactions with digital games, PX research has to consider aspects of interaction unique to the experience of play and the characteristics of digital games as a particular kind of product. Consequently, certain constructs are of increased importance to PX compared to classical UX, such as flow (Csikszentmihalyi, 1990), immersion (Jennett et al., 2008), or game enjoyment (Mekler et al., 2014). Furthermore, games are most often interacted with in a recreational context, unlike many products investigated in usability and UX research, such as productivity apps that serve a primarily functional purpose (Nacke & Drachen, 2011). Hence, theories and models are needed to explain digital game experiences, either developed explicitly for PX or adapted from related fields, such as psychology (Wiemeyer et al., 2016).

For the empirical study of digital game experiences, researchers employ various interdisciplinary research methods taken from fields such as HCI, computer science, neuroscience, and psychology (Nacke & Drachen, 2011). While PX measures cover numerous levels of experience, from behavioral over physiological to subjective measures (Wiemeyer et al., 2016), survey scales are among the most essential methods (Brühlmann & Mekler, 2018), likewise to UX research. Although some scales from UX research see use in the field of PX, many dedicated measures are designed specifically for the context of digital games. However, studies investigating the survey scales used in PX research have raised concerns regarding the psychometric quality of these measures (e.g., Kayser et al., 2021; Law et al., 2018; Memeti et al., 2022; von Felten et al., 2022). Thus, there is a need for robust and validated measures to study PX. One promising scale, the Player Experience Inventory (PXI), was recently proposed by Abeele et al.

(2020). The PXI is a 30-item scale measuring five functional consequences (ease of control, challenge, progress feedback, goals and rules, audiovisual appeal) and five psychosocial consequences (meaning, immersion, mastery, curiosity, autonomy) of interacting with video games. Alongside these ten constructs, the PXI's authors developed three items to measure enjoyment, recommended to be employed alongside the scale. The original work on the PXI, consisting of input from 64 game user researchers and five studies with a combined sample of 529 players, delivered favorable results concerning the PXI's psychometric quality (Abeele et al., 2020). However, the scale's quality had yet to be independently investigated, and the samples used to develop and validate the PXI were limited in terms of size and demographic diversity, motivating this thesis' fourth manuscript (Perrig, Scharowski, Brühlmann, et al., 2023).

The remainder of this thesis is structured as follows: First, the four manuscripts constituting this thesis are summarized, presenting each manuscript's motivation and aim, method, results, and a discussion based on each manuscript's findings. Afterward, overarching findings and insights from the four manuscripts and other publications and contributions related to this thesis are discussed concerning this thesis' overall topic of measuring and understanding UX through survey-scale-based research.

Summary of the manuscripts

The following manuscripts constitute this thesis. The first three manuscripts have already been published, whereas the fourth is under review.

1. **Perrig, S. A. C.**, Ueffing, D., Opwis, K., & Brühlmann, F. (2023). Smartphone app aesthetics influence users' experience and performance. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1113842

2. **Perrig, S. A. C.**, von Felten, N., Honda, M., Opwis, K., & Brühlmann, F. (2023). Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-pos. *International Journal of Human–Computer Interaction.* https://doi.org/10.1080/10447318.2023.2258634

3. Aeschbach, L. F., **Perrig, S. A. C.**, Weder, L., Opwis, K., & Brühlmann, F. (2021). Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY). https://doi.org/10.1145/3474660

4. **Perrig, S. A. C.**, Scharowski, N., Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2023). Independent validation of the Player Experience Inventory: Findings from a large set of video game players. *Manuscript submitted for publication.*

The following publications and contributions are related to this thesis but were omitted for the sake of brevity and focus. However, some are referenced in this thesis.

- Brühlmann, F., Memeti, Z., Aeschbach, L. F., **Perrig, S. A. C.**, & Opwis, K. (2024). The effectiveness of warning statements in reducing careless responding in crowdsourced online surveys. *Behavior Research Methods.*

- Kayser, D., **Perrig, S. A. C.**, & Brühlmann, F. (2021). Measuring players' experience of need satisfaction in digital games: An analysis of the factor structure of the UPEQ. *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, 158–162. https://doi.org/10.1145/3450337.3483499

- Memeti, Z., Brühlmann, F., & **Perrig, S. A. C.** (2022). LoL, why do you even play? Validating the motives for online gaming questionnaire in the Context of League of Legends. *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, 81-86. https://doi.org/10.1145/3505270.3558350

- **Perrig, S. A. C.**, Aeschbach, L. F., Scharowski, N., von Felten, N., Opwis, K., & Brühlmann, F. (2022). Measurement practices in UX research: A systematic quantitative literature review. https://doi.org/10.31234/osf.io/3jz67

- **Perrig, S. A. C.**, Scharowski, N., & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544549.3585808

- Scharowski, N., & **Perrig, S. A. C.** (2023). Distrust in (X)AI – Measurement artifact or distinct construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams.* https://doi.org/10.48550/arXiv.2303.16495

- Scharowski, N., **Perrig, S. A. C.**, Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, 5. https://doi.org/10.3389/fcomp.2023.1151150

- Scharowski, N., **Perrig, S. A. C.**, Aeschbach, L. F., von Felten, N., Opwis, K., Wintersberger, P., & Brühlmann, F. (2023). To trust or distrust trust measures: Validating questionnaires for rust in AI. *Manuscript submitted for publication.*

- Scharowski, N., **Perrig, S. A. C.**, von Felten, N., & Brühlmann, F. (2022). Trust and reliance in XAI – Distinguishing between attitudinal and behavioral measures. *CHI 2022 TRAIT Workshop on Trust and Reliance in AI-Human Teams.* https://doi.org/10.48550/arXiv.2203.12318

- von Felten, N., Brühlmann, F., & **Perrig, S. A.** (2022). Independent validation of the Video Game Dispositional Flow Scale with League of Legends players. *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, 44-50. https://doi.org/10.1145/3505270.3558351

First manuscript: "Smartphone app aesthetics influence users' experience and performance"

Motivation and aim of the study. In this study, we investigated how the aesthetics of a smartphone user interface impact the users' subjective experiences as well as their performance. While much previous research has looked at how aesthetics impacts user performance (see Thielsch et al., 2019, for an overview), little research to date has considered the context of modern smartphone devices. Hence, an online study was conducted to investigate how a manipulation of aesthetics, resulting in different variants of a smartphone interface, would affect the users' task performance and their subjective ratings of the interface's aesthetics and usability.

Method. The online study featured a between-subjects design with two conditions of manipulated interface aesthetics (high vs. low). A fictional event agency's interactive smartphone web application was developed and manipulated in terms of aesthetics to create two aesthetically different variants of an otherwise identical interface. The two variants differed in terms of symmetry, colorfulness, and complexity, while their functionality was kept as comparable as possible (e.g., not manipulating information architecture or page response time). An initial interface was developed based on inputs from four user interface and UX designers and then manipulated regarding aesthetics to form seven different variants. In a preliminary evaluation, 12 HCI researchers then rated these variants to select the two variants with the highest and lowest aesthetics for use as stimuli in the online experiment. A sample of 281 participants located in the United States of America was recruited over Amazon Mechanical Turk to participate in the online study. First, the survey platform automatically checked that participants accessed the study using a smartphone device. After providing informed consent and demographic data, participants were assigned to one of two conditions, either interacting with the stimuli's high- or low-aesthetics variant. They were asked to browse the stimuli variant, being told that they would have to answer a series of questions about the stimuli's content afterward, for which conscientious exploration was necessary. After the interaction, participants responded to six content-related questions, which were used to calculate a performance score. Each question asked for a specific detail about the fictitious company (e.g., "Since when has the Master Events agency been in business?"), with four possible answers, one of which was correct. Participants were given a point for each correct answer, resulting in a score ranging from 0 to 6. In addition, the time taken to respond to the content questions was tracked (i.e., the performance time). Performance was thus operationalized using two indicators: a performance score calculated from the number of correct answers to the content-related questions and the time taken to respond to these questions. Hence, high performance meant that a participant answered many questions from the information foraging task correctly while finishing the task as quickly as possible. After the performance questions, participants responded to two self-reported survey scales concerning their subjective perception of the stimuli: the 18-item English version of the VisAWI (Moshagen & Thielsch, 2010) to measure subjective aesthetics and the four-item Usability Metric for User Experience (UMUX, Finstad, 2010) to assess usability.

Results. Findings showed significant differences for both subjective measures, with higher ratings of perceived usability ($d = 0.86$) and visual aesthetics ($d = 1.26$) for the aesthetic compared to the unaesthetic variant (see Table 1). In contrast, no significant differences were found for the performance measures (i.e., performance score and time). However, additional statistical procedures, namely equivalence tests, bootstrapping, and descriptive statistics, suggested that aesthetics impacted the performance score ($d = 0.22$) but that this effect was non-significant due to methodological reasons (e.g., ceiling effect for performance score, sample size). In addition, the quality of the two subjective survey scales employed (i.e., UMUX and VisAWI) was investigated prior to interpreting the data, using internal consistency coefficients, CFA, and EFA. This was done because neither scale was developed for the mobile device context, and the VisAWI was validated only in German while we used the English version. Results here suggested that the scales' structure was distorted by using both positively and negatively formulated items, an issue revisited in the second manuscript.

Table 1

Descriptive values for key variables in the first manuscript sorted by stimuli variant.

	Aesthetic (n = 139)			**Unaesthetic (n = 142)**		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
VisAWI - Total Score	5.62	0.95	2.94 - 7.00	4.00	1.54	1.23 - 7.00
UMUX Score	80.19	18.47	25.00 - 100.00	61.44	24.43	4.17 - 100.00
Performance Time (minutes)	2.52	2.03	0.22 - 14.07	2.65	2.27	0.28 - 13.32
Performance Score	5.26	1.23	0.00 - 6.00	4.95	1.58	0.00 - 6.00

Discussion and conclusion. The present work showed that smartphone interface aesthetics positively impact users' subjective experiences, which is in line with past findings from other device contexts (e.g., Minge & Thüring, 2018; Sonderegger & Sauer, 2010). While the significant difference in subjective aesthetics was also taken as evidence for the success of the stimuli manipulation, results concerning subjective usability speak for a halo effect (Thorndike, 1920) of interface aesthetics on perceived usability, as previously suggested by Tractinsky et al. (2000). Furthermore, results pointed towards an effect of aesthetics on objective user performance, although no significant differences were found. Methodological challenges encountered during the research possibly explain this absence of significant differences. In particular, both groups had relatively high mean performance scores, and most participants were able to correctly respond to the majority of questions, suggesting a ceiling effect (i.e., responses clustered near the highest possible score). Findings thus highlight that the choice of measure for the users' performance critically influenced the results. Given the highly contextual nature of possible performance tasks, standardized scales such as those used to measure subjective aesthetics and usability are not realistic for drawing inferences about performance in most cases. Consequently, researchers must carefully think about how they want to operationalize performance and critically investigate the quality of their performance measures while combining more than one performance indicator to capture different aspects of user interaction.

In addition, the present study demonstrated that scales previously validated in research might not hold up when used in a new context or with another target population, highlighting that scale validation is an ongoing process not limited to a single study or paper. Both the UMUX used to measure usability and the VisAWI aesthetics scale were not initially developed for a mobile device context, leading us to re-investigate their psychometric quality. In this regard, results showed that the item tone (positive or negative formulation) distorted the factor structure of both scales, something also noted for other scales in HCI (e.g., Lewis & Sauro, 2017; Perrig, Scharowski, & Brühlmann, 2023). These results directly inspired the next manuscript.

Second manuscript: "Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-pos"

Motivation and aim of the study. Given the findings from the first manuscript concerning a distorted scale structure for the VisAWI, this work sought to investigate the effect of negative or positive item formulation on the quality of the survey scale. Furthermore, the work aimed at independently validating the English version of the VisAWI and its short version, the VisAWI-S, given that past work developing and validating the scale was conducted for the German version. The VisAWI consists of 18 items, measuring four facets of aesthetics: simplicity, diversity, colorfulness, and craftsmanship. The scale contains two negatively formulated items for each of the four facets. Negatively formulated or reverse-coded items are formulated opposite to the direction of the other scale items, which is supposed to counteract unwanted response biases, such as acquiescence or extreme response (Sauro & Lewis, 2011). However, numerous research articles have demonstrated that using negative item formulation results in more issues than the biases they are designed to prevent (Sauro & Lewis, 2011). It might thus be better to avoid using such items, motivating the development of an alternative VisAWI version, which only consists of positively formulated items.

Item development. In an initial step, alternative versions for each negatively formulated VisAWI item were developed. Two authors independently drafted a first set of items by removing negations or searching for suitable antonyms, which were then discussed and combined in a second step. Next, items were reviewed by an English language expert, who discussed the items with the first author to settle on an initial set of alternatives. The resulting set of 18 items, two to three per negative VisAWI item, was then investigated in a first online study.

Method study 1. A between-subjects online study was conducted to reduce the set of positive alternatives, ideally to one alternative per negatively formulated VisAWI item. The same stimuli already used in the first manuscript were employed for this study, but this time in a desktop setting. Participants recruited over Prolific ($N = 41$) were randomly assigned to one of the two stimuli sites and asked to browse it in order to respond to the performance questions developed for the first manuscript. Afterward, participants responded to all 18 items of the VisAWI and all 18 positive alternatives.

Results study 1. We first considered descriptive statistics, item difficulty, item variance, discriminatory power, and inter-item correlations. However, no items were removed or flagged as suspicious based on these analyses. Thus, correlations between the alternative and original items were calculated to search for the most appropriate alternatives. Based on these results, combined with the English language expert's input from the item development, an initial version of the VisAWI-pos was created. The VisAWI-pos was then compared to the original VisAWI based on descriptive statistics, internal consistency, the capability to distinguish the two stimuli websites, and correlations. Descriptive statistics for the two scale versions were comparable, and reliability for both scales was excellent. Furthermore, the aesthetic stimuli variant was rated higher in both scale versions, and scores for the two scales correlated highly, both overall and per facet. Thus, we concluded that the newly developed VisAWI-pos was a viable alternative to the original VisAWI, although additional research with a larger sample and different stimuli was needed.

Method study 2. To investigate the psychometric quality of the VisAWI-pos, comparing it to the English VisAWI and VisAWI-S, a pre-registered between-subjects online study was conducted. Using Prolific, 966 viable participant responses were collected. As stimuli, a set of 12 websites covering six content areas (arts and entertainment, law & government, news & media publishers, science & education, food & drink, and lifestyle) was prepared, with one popular and one unpopular website per content area (based on rankings from similarweb.com). For each website, two content questions were developed to ensure that participants would actually interact with the sites. After interacting with the website, participants responded to all VisAWI and VisAWI-pos items, a scale on visual aesthetics (Lavie & Tractinsky, 2004), the User

Experience Questionnaire (Laugwitz et al., 2008), the positive version of the System Usability Scale (Brooke, 1996; Sauro & Lewis, 2011), the Web-CLIC-S (Thielsch & Hirschfeld, 2021), and the single-item Net Promoter Score (NPS, Reichheld, 2003).

Results study 2. Descriptive statistics, item difficulty, item variance, discriminatory power, and inter-item correlations showed that while the items for the original VisAWI and the VisAWI-pos sometimes differed, none exhibited problematic values. Reliability indicators, namely coefficients alpha (Cronbach, 1951) and omega (McDonald, 1999), showed very high internal consistency for all three VisAWI versions (original, positive, and short), with equal or sometimes higher values for the VisAWI-pos compared to the other two. Next, CFAs were conducted to investigate the scale versions' model fit, which was judged following recommendations by Hu and Bentler (1999). Given that the multivariate normality was violated, a robust maximum likelihood estimator was used. Results, presented in Table 2, suggested a good model fit for the VisAWI-S but slightly less for the VisAWI-pos and especially for the original VisAWI.

Table 2
Fit indices for CFA models of the VisAWI versions in study 2 of the second manuscript.

Model	χ^2	df	p-value χ^2	RMSEA	SRMR	CFI
VisAWI	792.30	131	$< .001$	.083	.047	.933
VisAWI-pos	576.10	131	$< .001$	.067	.047	.960
VisAWI-S	13.33	2	$< .01$	.091	.017	.991

Note: Robust values are reported wherever possible.

Given the somewhat sub-optimal CFA model fits for the VisAWI and VisAWI-pos, we continued with multiple EFAs to search for ways to improve model fit. Reporting of the analysis focused on four-factor solutions, following the proposed theoretical model behind the VisAWI, while alternative solutions were explored and provided in the supplementary materials. For the VisAWI, a four-factor EFA explained 66% of cumulative variance but exhibited problematic loadings for six items, of which four were negatively formulated. In particular, three out of four craftsmanship items did not load onto a dedicated factor but rather on the same factor as items for simplicity and diversity. It was thus concluded that the negative items are a likely reason for the sub-optimal CFA results of the original VisAWI but that a four-factor solution was still most suitable for the scale. In addition, the craftsmanship items did not form a distinct factor, pointing to more substantial issues with this facet. Concerning the VisAWI-pos, a four-factor EFA explained 69% of cumulative variance, with items for simplicity, diversity, and colorfulness loading well onto three separate factors. For craftsmanship, however, items did not load well onto a distinct factor, either exhibiting high cross-loadings or loading onto the same factor as the simplicity items. It was thus concluded that a four-factor solution was reasonable for the VisAWI-pos but that there also are issues with the craftsmanship items.

Next, convergent and discriminant validity was assessed using correlations between the VisAWI versions and related scales. Results mostly were as expected, supporting the convergent and discriminant validity for all three scale versions. Finally, the VisAWI's ability to differentiate between the 12 stimuli websites was investigated as evidence of criterion-related validity. Results were comparable for all three versions and showed that the scales' ratings differed significantly between the websites. Furthermore, results for the three VisAWI versions were similar regarding which websites received higher or lower scores (i.e., rankings of the websites based on means) as well as which websites' ratings fell above or below the VisAWI cutoff of $\overline{x} \geq 4.50$ for aesthetic websites suggested by Hirschfeld and Thielsch (2015).

Discussion and conclusion. The present work showed that replacing the negative items of a scale with positive alternatives, in this case for the VisAWI, can improve the scale's psychometric qualities. While results suggested that the VisAWI-pos still holds room for improvement regarding the craftsmanship items, they also displayed that the remaining three facets of the scale were more pronounced in the new version. In contrast, the negative items were shown to cause issues for the original VisAWI, which was to be expected based on past research on reverse-coded items (e.g., Salazar, 2015). However, findings also showed that the psychometric quality of the three VisAWI versions was comparable to results previously reported for other language versions of the scale, namely in Arabic (Abbas et al., 2022), German (Moshagen & Thielsch, 2010), and Farsi (Saremi et al., 2023). Nevertheless, given the reported findings, the VisAWI-pos was considered to be a preferable alternative to the original VisAWI. Concerning the issues identified for the craftsmanship facet, three possible reasons were discussed. First, there has been a homogenization of website design in the last years, with page layout, in particular, becoming more similar (Goree et al., 2021). Therefore, what is or is not considered a well-crafted website has likely changed since the VisAWI was initially developed. Second, multiple results suggested that craftsmanship might be an overarching construct rather than a facet of aesthetics distinct from the other three VisAWI facets, raising questions about the adequacy of the scale's theoretical model. Finally, participants' cultural backgrounds possibly influenced the craftsmanship ratings, given that no issues were reported for the German-speaking samples used to develop the VisAWI but for Arabic- and English-speaking participants. Thus, the cultural measurement invariance of the VisAWI requires further research.

In summary, the paper provided an improved way to measure aesthetics based on the promising VisAWI scale, which has the added benefit of avoiding issues related to negatively formulated items. The study further demonstrated how switching from negative to positive item wording can favorably influence the psychometric quality of a survey scale. It also shows that a scale's quality might change over time and depending on the cultural setting in which it is applied.

Third manuscript: "Transparency in measurement reporting: A systematic literature review of CHI PLAY"

Motivation and aim of the study. This third manuscript looked at how survey scales are currently used and reported on in HCI research, focusing on the particular area of PX research (UX "in the specific context of digital games" (Nacke & Drachen, 2011, p. 1)). Given the popularity of survey scales to collect data in PX research (Brühlmann & Mekler, 2018), the study aimed to determine how transparently measurement based on survey scales is currently reported in the research literature. In particular, the study looked at whether researchers report *what*, *how*, and *why* constructs are measured using survey scales. Transparency of measurement reporting is vital because only with transparent reporting can reviewers and readers of a paper form an opinion about the rigor and quality of a study's methods and, consequently, its results (Flake & Fried, 2020). Furthermore, transparent reporting is essential for accumulating research knowledge across studies and for developing and refining theories. However, to date, little to no research in HCI has investigated the transparency of measurement reporting in scale-based research, neither for UX nor for PX.

Method. A systematic literature review of PX research employing survey scales was conducted to investigate the current state of transparency in measurement reporting. The review specifically focused on proceeding to the ACM's Annual Symposium on Computer-Human Interaction in Play (CHI PLAY) for the year 2020. All full papers published at CHI PLAY 2020 were screened for inclusion in the sample ($n = 48$). A codebook for gathering data on the relevant variables was developed based on questions for transparent measurement reporting by Flake and Fried (2020) and refined by the paper's first author based on a random subset of the eligible papers. Afterward, subsets of the 48 papers were coded by the first three authors of the paper, with overlap between the subsets for the calculation of inter-rater agreement. Codings covered the topics of construct definition, construct operationalization, measurement selection, modification of measurements, and self-development of measurements.

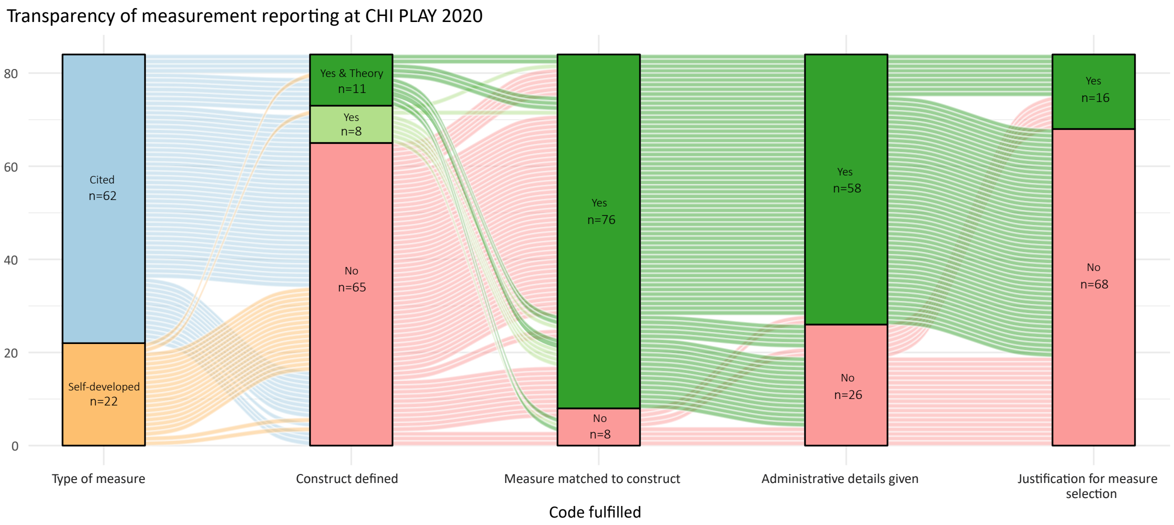


Figure 1

Alluvial plot summarizing key findings from the third manuscript on measurement reporting at ACM CHI PLAY 2020 for all instances of measurement ($N = 84$).

Results. Key findings of the literature review are summarized in Figure 1. Out of 48 full papers screened, 24 reported employing at least one survey scale to measure at least one construct and were thus eligible for inclusion. Across the 24 eligible papers, a total of 84 instances of measurement were recorded, of which the majority ($n = 62$) were done using cited measures (i.e., survey scales for which a source was provided), which is depicted in the first bar of Figure 1. A total of 67 different constructs were measured, most only once throughout the papers ($n = 60$). Furthermore, 41 different cited measures were identified, contrasting 22 self-developed measures, while half of all 24 eligible papers contained at least one self-developed measure.

Results showed that only around a fifth of all measured constructs were defined (22.62%), with even fever specified within a theory (13.10%, see second bar in Figure 1). Half of all papers defined at least one measured construct, but only two defined all of them. Concerning operationalization, most authors matched the constructs to the measures used (90.48%, third bar in Figure 1) while details on the administration of the measures (e.g., Likert-type scale used) were also provided by the majority (69.05%, fourth bar in Figure 1). In contrast, justification for measurement selection was rare (19.05%, fifth bar in Figure 1), while no justification for developing a new measure was given in any of the cases employing self-developed measures. In addition, results showed that almost half of all cited measures were modified (38.71%). Justification for these modifications was frequently neglected, but examples were reported in almost all cases. Finally, coding of the sources provided for the 41 cited measures showed that almost half (46.34%) did not contain any evidence of construct validity (i.e., EFA or CFA), although additional sources containing evidence of validity could be identified among the citations for 30 of the 41 cited measures (73.17%).

Discussion and conclusion. Findings simultaneously highlighted strengths and weaknesses regarding the current state of transparency in measurement reporting at CHI PLAY. Concerning *how* constructs are measured, most researchers were transparent in their reporting. However, results displayed a need for more transparency concerning *what* is being measured (i.e., the definition of constructs) and that reporting is limited in terms of theoretical considerations and the embedding of survey scales and their constructs within a theory. The displayed lack of transparency makes it hard to judge the validity of scales and their usage, as well as the validity of the reported results and the studies as a whole (Flake & Fried, 2020). Nontransparent reporting further hinders the accumulation of research knowledge, such as through meta-analyses, as well as the formation and refinement of theory. This is especially problematic when considering that without such theory, statistical validation procedures lose their meaning (Maul, 2017). Results further demonstrated that researchers report little on *why* a chosen construct was measured the way it was, indicating a lack of rigor concerning the process of measurement selection. In addition, results revealed that researchers rely on self-developed measures for constructs where validated questionnaires would exist without providing reasoning for doing so. This further questions the validity of results derived from the measurement, complicating the accumulation of research knowledge. A measurement selection model was derived from the findings to improve the reporting quality of scale-based research. The model consists of three steps and guides researchers through the measurement selection process, helping them to consider and report what they were measuring, including clear definitions of target constructs, how they were measuring (i.e., how the target constructs were operationalized), and why the researchers chose to measure in the way they did (i.e., reporting evidence of validity). Overall, the paper showed a clear need for more transparent reporting in scale-based PX research while providing actionable recommendations for future research to become more transparent. While the present work only focused on PX research, and in particular on one year of CHI PLAY proceedings, comparable findings were also found in a review (preprint) of UX research spanning four years of proceedings to the ACM Conference on Human Factors in Computing Systems (Perrig et al., 2022). Consequently, it is reasonable to assume that results also apply to other areas of HCI research beyond the study of PX.

Fourth manuscript: "Independent validation of the Player Experience Inventory: Findings from a large set of video game players"

Motivation and aim of the study. Motivated by the need for adequately validated measurement tools to study PX, the fourth manuscript set out to independently validate the Player Experience Inventory (PXI, Abeele et al., 2020), a scale for the measurement of functional and psychosocial consequences of interacting with video games. Functional consequences are "the immediate and tangible consequences that are experienced directly by consumers, during the use of the product" while psychosocial consequences "exceed the immediate usage level and reach into the social or, psychological level" (Abeele et al., 2020, p. 3-4). The PXI consists of 30 items, measuring ten constructs: five for the functional consequences (ease of control, challenge, progress feedback, goals and rules, and audiovisual appeal) and five for the psychosocial consequences (meaning, immersion, mastery, curiosity, and autonomy). In addition, the original authors developed three items for enjoyment, which are not officially part of the PXI but suggested to be used alongside it. While initial results on the psychometric quality of the PXI were promising, independent validation was needed for several reasons. First, a literature review of articles citing papers on the PXI ($N = 45$) conducted as preparation for the validation study showed discrepancies between the theoretical model of the PXI and the way researchers employed the scale. In particular, multiple papers reported on calculating an overall PX score, although the original authors never suggested such a procedure. Furthermore, many researchers employed the enjoyment items, whose psychometric quality had yet to be investigated. In addition, scale validation is an ongoing process (Furr, 2011), which the PXI's authors acknowledged themselves. Given that participants in past PXI samples were predominantly young men, validation with other populations was needed. Additionally, other studies re-investigating the psychometric quality of PX scales have shown that some scales do not always hold up under new conditions, at least not without modification (e.g., Kayser et al., 2021; Law et al., 2018; Memeti et al., 2022; von Felten et al., 2022). We thus set out to independently validate the PXI in a large-sample pre-registered online study.

Method. In the online study, 1518 participants recruited over Prolific were asked to recall a digital game they recently played or know well. They rated their experience with the game using the PXI and related standardized survey scales, namely the Player Experience of Need Satisfaction scale (PENS, Ryan et al., 2006), the AttrakDiff (Hassenzahl, 2004), and the interest/enjoyment subscale from the Intrinsic Motivation Inventory (IMI, Ryan & Deci, 2000; Ryan et al., 1983). For recalling the recent or memorable gaming experience, we asked participants to describe the game in at least 50 words, following the critical incident technique common in HCI research (e.g., Bopp et al., 2016; Seckler et al., 2015).

Results. The analysis focused on different forms of psychometric quality investigation for the PXI, mainly following the process reported in the initial work on the PXI. Item analysis, consisting of descriptive statistics, item difficulty and variance, discriminatory power, and iter-item correlations, was mostly inconspicuous, with a few exceptions which were kept in mind for further analyses. In the next step, multiple CFAs were performed to investigate how various theoretical models would fit the collected PXI data. The theoretical models primarily differed in whether they contained higher-order factors, either for the functional and psychosocial consequences, for PX, or for both. In addition, a model including the enjoyment items and a respective factor for enjoyment was also investigated. A robust maximum likelihood estimator was used for all CFAs because multivariate normality was not given. Results from these CFAs are presented in Table 3. Findings mainly favored those models without higher-order factors, showing less optimal fits for those models with higher factors for PX and/or the consequences. Furthermore, results demonstrated that including the enjoyment items had no adverse effect on model fit.

In addition to the CFAs, coefficients of internal consistency (i.e., alpha (Cronbach, 1951) and omega (McDonald, 1999)) were calculated as indicators of the PXI's reliability, which favored both the overall PXI, with and without enjoyment, and its subscales. The only exception was immersion, which fell just below the desired threshold. To investigate the PXI's convergent and discriminant validity, values of

Table 3

Fit indices for CFA models of the PXI in the fourth manuscript. Models 1 and 5 were assessed without higher-order factors, and models 2, 3 & 4 included varying higher-order factors.

Tested model	χ^2	**df**	**p-value** χ^2	**RMSEA**	**SRMR**	**CFI**	**TLI**	χ^2/df
1) 10 factors (original PXI)	1053.213	360	$< .001$	.041	.041	.956	.946	2.926
2) 10 factors + 2 factors consequences	1866.132	394	$< .001$	.057	.079	.906	.896	4.736
3) 10 factors + 1 factor PX	2000.001	395	$< .001$	.060	.079	.897	.886	5.063
4) 10 factors + 2 factors consequences + 1 factor PX	1861.395	393	$< .001$	.057	.079	.906	.896	4.736
5) 11 factors, incl. enjoyment	1278.195	440	$< .001$	.041	.040	.955	.946	2.905

Note: Robust values are reported wherever possible.

composite reliability (CR), maximum shared variance (MSV), and average variance extracted (AVE) were calculated based on the 11-factor CFA. Results regarding CR favored the scale's reliability, while AVE values were good for most constructs except for mastery, immersion, challenge, and ease of control, indicating good convergent validity. Regarding discriminant validity, comparing AVE to MSV values favored all constructs except for immersion. At the same time, comparing the constructs' square root of the AVE to the inter-construct correlations also supported all constructs' discriminant validity except for immersion.

Finally, criterion validity was assessed in two ways: first, by considering correlations between the PXI's constructs and the related scales, and second, by calculating a hybrid structural equation model testing the theoretical relationship between the psychosocial consequences, the functional consequences, and the enjoyment items. The correlations' results were mainly as expected, with strong positive correlations between most of the PXI's constructs and their counterparts from the other scales. The only exceptions were moderate correlations for challenge and meaning and weak correlations for progress feedback and goals and rules with their respective counterparts. Regarding the hybrid structural equation model, all relationships were as expected and similar to those reported in the original work on the PXI, thus further supporting the PXI's criterion validity and its theoretical model.

Discussion and conclusion. Overall, results from the study showed that the PXI is of good psychometric quality, considering typical reliability and validity indicators. Results were also comparable to those initially reported for the PXI (Abeele et al.,

2020) and those presented for a German version of the scale (Graf et al., 2022). In addition, the crowdsourced sample used in the present study differed from the samples used in past work on the PXI, mainly in terms of more balance concerning age and gender. Findings thus demonstrated that the PXI maintains its psychometric quality when used with other participant populations. The independent validation further confirmed that a theoretical ten-factor model for the PXI, without higher-order factors, best suits the scale while demonstrating that an 11-factor model, including the enjoyment items, likewise performs well.

However, results also showed issues with the PXI, mainly concerning the construct of immersion and its differentiation from the other constructs of the scale. While immersion is a popular construct to be measured and studied with the PXI, it has been criticized in past research (e.g., Aeschbach et al., 2022). Thus, more theoretical work is needed regarding a more consistent definition of the immersion construct and to distinguish it from related constructs. Finally, the present results gave little support for an overall PX score. Beyond showing what should and should not be measured using the PXI, this further raised the question of how sensible calculating an overall score for experience is in the first place. In most cases, interactive media, such as digital games, might not address all constructs measured with the PXI equally, and a particular game is presumably also not designed to do so. The resulting ratings for the PXI are high for some constructs and low for others. When calculating an overall PX score, these differences between the constructs are averaged out and thus neglected, leading to a low score for a product, even if it scores high on the relevant constructs for which it was designed. For example, a particular game might have been designed for a high difficulty level while not trying to provide ease of control. In this case, a low score in ease of control does not mean there is a design problem to be fixed but instead reflects the design intentions. Therefore, the PXI's strength lies not in providing an overall product assessment but in comparing different digital games or versions of the same game regarding the individual constructs.

General discussion

The goal of this thesis was to understand better how UX is currently measured in research using survey scales, what this reveals about researchers' understanding of UX, and how scale-based UX research can be improved.

In this regard, the first manuscript has shown how methods and results from one form of product (e.g., desktop websites) can be translated to a new setting (e.g., mobile context) and where they are challenging to transfer. While some previous research results, such as the relationship between perceived aesthetics and usability, were consistent when translated into the new context, other results were not (e.g., the psychometric quality of the VisAWI). The first manuscript has also shown how difficult it can be to measure certain constructs (e.g., performance) and how important it is to choose suitable methods, evident in that the choice of performance measure likely influenced the obtained results.

The first and second manuscripts also tackled the central construct of aesthetics, showing how challenging it is to define and measure. Past research suggested that contradictory results on the effects of usability and aesthetics are likely due to the use of diverse measurement methods (Hassenzahl & Monk, 2010), thus highlighting the importance of adequate survey scales for aesthetics. While the VisAWI is a promising way to measure aesthetics, the presented work has also shown that it comes with certain limitations. Namely, the first manuscript showed that the scale exhibits a two-factor structure instead of the initially proposed four factors for the scale. Based on these findings, the second manuscript investigated one possible reason for this distorted factor structure: negatively formulated items. Through the development of an alternative VisAWI version with positive items only, the VisAWI-pos, we were able to demonstrate that the replacement of the reverse-coded items with positive alternatives improved scale quality and model fit. This finding is relevant not only for researchers who employ the VisAWI but also for other HCI scales, given that they were also shown to be affected by the inclusion of reverse-coded items (e.g., Lewis et al., 2013; Perrig, Scharowski, & Brühlmann, 2023; Sauro & Lewis, 2011).

Furthermore, both the VisAWI and VisAWI-pos exhibited issues beyond those linked to negative item formulation, namely, with the "craftsmanship" construct. Through the results on the problematic craftsmanship facet of the VisAWI, the second manuscript demonstrated how the quality of a scale can change over time. For example, a scale developed a decade ago might not be suitable anymore for current websites (e.g., because of changes in what is considered a well-crafted site) and thus has to be re-evaluated and adapted to fit new designs. Another explanation could be that a scale's quality changes depending on the cultural or personal background of the participants, an explanation also considered in the second and fourth manuscripts. However, both of these arguments speak to the importance of continuous quality re-investigation, which is addressed in more detail below.

In addition, the third manuscript has shown that the quality and proper use of scales are rarely given the attention they deserve. In general, the investigation into current practices of scale-based HCI research has shown that there still is much room for improvement concerning transparent measurement reporting both in PX (Aeschbach et al., 2021) and UX (Perrig et al., 2022). This not only makes a comparison of research findings across studies complex or even impossible but also shows the importance of proper training and care when it comes to using survey scales for measurement, a point previously made in Law et al. (2014). Thus, researchers need to use survey scales properly to reap their full potential, which the third manuscript addressed through a measurement selection model.

Finally, the fourth manuscript focused on a specific scale for the measurement of constructs central to PX, exemplifying how measurement is conducted there. Results from the independent re-investigation showed that the scale performed well. Findings also pointed to the importance of theory and unambiguous construct definition, illustrated in the confusion among researchers about which theoretical model to use for the scale, or rather how to evaluate the scale's responses, as well as the sub-optimal results regarding the somewhat controversial construct of immersion (e.g., Aeschbach et al., 2022). Past work has already criticized how essential terms in PX often have no agreed-upon definition and are used interchangeably (Nacke & Drachen, 2011), which appears to also apply to immersion. Furthermore, the fourth manuscript has demonstrated that there is little sense in measuring an overall PX score, at least not with the PXI. Given that this might also apply to UX, part of the discussion below will debate why it might be meaningless to measure overall UX as a single score, given the nature of UX as an umbrella construct.

In summary, the four core manuscripts of this thesis highlighted three areas of research regarding how to measure and understand UX. First, the remainder of this discussion will reflect on the importance of theory as a foundation for measurement. Second, the discussion will address how scale validation is an ongoing process. Finally, the discussion will consider what the present results tell us about how UX should be measured and understood before limitations and opportunities for future research are presented.

The importance of theory as a foundation for measurement

Overall, the manuscripts constituting this thesis have emphasized the importance of theory as a foundation for measurement. While the first and second manuscripts have shown how a scale's factorial structure can deviate from the expected theoretical basis due to methodological reasons (e.g., negative items), the second manuscript also showed how the suitability of a scale's theoretical model might change over time, possibly due to changes in product design and perception. Likewise, the second and fourth manuscripts have shown how certain constructs within a scale can be difficult to correctly distinguish from others, namely craftsmanship in the case of the VisAWI (second manuscript) and immersion for the PXI (fourth manuscript). Meanwhile, the third manuscript has highlighted how researchers need to give more attention to the process of measurement selection while also considering a scale's theoretical foundations, such as construct definitions or specification of a survey scale's constructs within a theory. Simply because a scale performs well or seems to perform well does not necessarily mean it measures the construct it is intended to measure, as demonstrated in past research (Maul, 2017). Hence, researchers must carefully consider the theoretical foundations of the scales they use. Past research has already raised concerns regarding missing theory in UX research (Law et al., 2014). Thus, future work needs to pay more attention to the theoretical foundations of UX and its constructs, especially when developing and employing survey scales. However, a detailed look into current theories and future theoretical avenues for UX research was beyond the present thesis' scope.

Scale validation as an ongoing process

Numerous projects conducted as part of the present thesis have shown that scales often do not hold up under re-evaluation, or at least hold room for improvement requiring modification or refinement (e.g. Kayser et al., 2021; Memeti et al., 2022; Perrig, Scharowski, & Brühlmann, 2023; Perrig, Scharowski, Brühlmann, et al., 2023; Perrig, von Felten, et al., 2023; von Felten et al., 2022). In this regard, the second manuscript has shown, for example, how what is more or less influential for good UX can change over time. In particular, the manuscript demonstrated how conceptualizations of constructs central to UX, such as aesthetics, can vary with developments in design, as evident in the identified issues associated with the craftsmanship facet of the VisAWI. Such changes require researchers to re-evaluate and refine the items of a survey scale to ensure that it still measures what it was initially designed to measure.

Similarly, the first, second, and fourth manuscripts have further highlighted the importance of measurement invariance across different groups of participants and cultural backgrounds. For example, the first two manuscripts have displayed how a scale validated in one language (in this case, German) might not perform as well when translated to another language (e.g., English). Meanwhile, the first manuscript has also illustrated how there is a need to re-evaluate a scale when switching the context of use (e.g., from desktop to mobile websites), and the fourth manuscript has shown how crucial it is to study the quality of a scale with different and diverse populations of participants. Furthermore, the third manuscript has uncovered that survey scales are frequently adapted in PX research, possibly affecting their psychometric quality (Juniper, 2009) and that the quality of scales is often questionable, with sources cited for the measures not providing sufficient evidence of their validity.

Thus, rather than simply assuming that the scales they use are of adequate psychometric quality, researchers should ideally re-evaluate the psychometric quality of a scale, especially in a new context or when collecting data from a novel population (Furr, 2011), as was done in the first manuscript. Because such re-evaluation is not always reasonable (e.g., due to sample sizes or resource constraints), independent validation efforts, such as those reported in the second and fourth manuscripts, are critical to provide researchers with additional confidence in the quality of their measures and thus in the results collected with them.

Measuring and understanding UX

Overall, part of this thesis' goal was to understand better what UX is and how to define it by examining how UX is currently measured using survey scales. However, the research reported across the four manuscripts comprising this thesis has further shown the complexity of UX or, more precisely, the sub-fields on which this thesis focuses (i.e., aesthetics and PX).

Past research has shown that both in research on PX (Perrig, Scharowski, Brühlmann, et al., 2023) and UX (Bargas-Avila & Hornbæk, 2011; Perrig et al., 2022; Pettersson et al., 2018), researchers sometimes measure an overall construct representing the experience. For example, in Pettersson et al. (2018), "generic UX" was the most frequently measured dimension of UX, describing instances where authors of the examined papers understood "UX as a general construct and did not specify which aspects they studied in detail" (Pettersson et al., 2018, p. 5). However, the usefulness of attempting to develop a comprehensive understanding of a broad construct, such as UX, or measuring it as a general construct is debatable. Instead, it appears that there is more benefit in gauging the components that are important to an experience rather than trying to quantify the experience as a whole. This was exemplified by the results of the fourth manuscript on the PXI, where theoretical models incorporating factors for overall PX fit the data worse than those models just considering the individual constructs of the scale. Thus, trying to understand and break down the entirety of a person's experience with an interactive product into a single construct and one number does not do justice to the complexity of said experience. As past work has pointed out, UX can be understood as an umbrella construct, which is difficult to understand and define (Tractinsky, 2018). Hence, the goal for future research should be to properly understand and differentiate the individual components of the experience and the core constructs comprising UX rather than trying to reduce the entirety of UX into a single measurable construct.

As shown in the third manuscript concerning PX, researchers need more common ground on which constructs to measure and how to define them. Comparable findings were also found in a review (preprint) of UX research (Perrig et al., 2022). More attention needs to be placed on transparent reporting and the measurement process itself, which can be done by following the measurement selection model presented in the third manuscript. Furthermore, high-quality scales (i.e., validated and re-validated) need to be used to ensure that differences in results are due to the experimental factors of interest and not due to methodological issues, such as negative items or inconsistent use of dissimilar measures. Researchers should further follow calls from past research (e.g., Hornbæk, 2006; Hornbæk & Law, 2007; Sauro & Lewis, 2009) to use scales from prior research rather than relying on ad hoc scales or modified versions of standardized measures. In addition, conceptual work on UX is required to link individual research efforts by formulating models and theories that explain previously unrelated phenomena transpiring in interaction (Oulasvirta & Hornbæk, 2016). In the past, results by Law et al. (2014) have raised concerns regarding a lack of theory in UX research, and Law (2011) emphasized the importance of proper theories of UX to form meaning out of the collected data from qualitative and quantitative research. In this process, survey scales play a crucial role as one of multiple methods, helping researchers establish a more profound understanding of what UX is and what constructs are essential to UX.

In summary, human experiences formed from interacting with technology are a complex field of study comprising many interlocking parts. Hence, UX should not be understood as a singular construct but as consisting of many different parts forming the users' experiences. However, proper measurement practices are needed to study the essential elements that form UX adequately. As was shown in the third manuscript for PX and in other reviews on UX (Bargas-Avila & Hornbæk, 2011; Perrig et al., 2022; Pettersson et al., 2018), there is still room for improvement when it comes to how researchers employ survey scales in their work. Consequently, more training, appreciation, and care are needed to reap the full potential that survey scales as a method would bring.

Limitations and future directions

First, the present work composing this thesis focused on survey scales. While among the most popular methods in UX research, other approaches to investigating UX, such as qualitative methods, exist. Although equally important to the study of UX, these were outside the scope of the present thesis.

Second, the present work focused on academic research on UX. However, the term and concept of UX are also crucial in the industry, where perspectives and practices likely differ. Thus, systematically talking to practitioners or studying how they employ survey scales and which ones they use could further improve our understanding of how UX is measured and understood there. In addition, more profound knowledge of how UX is understood in the industry compared to academia could reduce translational barriers between the two, which is a crucial but difficult challenge in HCI (Colusso et al., 2019).

Third, the samples used in the present work were limited in multiple ways. The first, second, and fourth manuscripts were all based on data from crowdsourced samples. While past research has demonstrated these samples to be of high data quality and comparable to other more traditional populations such as students (Douglas et al., 2023), results might differ with other samples. Furthermore, we worked with English-speaking Western samples for all projects constituting this thesis. While using entirely Western participant samples is common in HCI research, it has also been a source of critique (Linxen et al., 2021). In addition, the present work relied on data collected in online studies. This allowed us to collect samples large enough for certain analyses, such as CFA, which was especially crucial for the scale validations reported in the second and fourth manuscripts. However, while online studies are a standard method able to deliver insightful and high-quality data if adequately used (for recommendations, see Brühlmann et al., 2020), results might differ for data collected in a lab setting or in field studies.

Fourth, the present work focused on aesthetics in the first two manuscripts and PX in the second two. Thus, this thesis considers only specific areas of UX, which might only be representative of some of UX research. Other areas of UX and HCI research remain to explore regarding the questions addressed in the present thesis. For example, research on survey scales designed for human-AI interaction has also identified problematic psychometric quality and difficulty conceptualizing and measuring core constructs such as trust and distrust in AI (Perrig, Scharowski, & Brühlmann, 2023; Scharowski & Perrig, 2023; Scharowski et al., 2023).

Finally, the present thesis has demonstrated the importance of theory for UX research and a deepened understanding of the individual constructs that make up UX. While a detailed examination of the current state of UX theory and the formulation of new theories where needed was beyond the scope of this thesis, previous research has already highlighted the need for more theoretical work in HCI (Oulasvirta & Hornbæk, 2016). Thus, more work in this direction is required to develop a more profound understanding of UX and its constructs. Data gathered through survey scales can be crucial to these efforts, but only if the scales are of sufficient psychometric quality.

Conclusion

Survey scales are among the most popular methods in UX research. However, they ought to be of adequate psychometric quality and be correctly used to fully harvest their potential. The present thesis presented four manuscripts examining survey scale usage in UX research from various perspectives. In the first manuscript, one central question of UX research, the effect of interface aesthetics on subjective aesthetics, subjective usability, and objective performance, was investigated in the novel context of smartphone devices. Results showed that interface aesthetics positively impact users' subjective experiences while pointing toward a positive impact on their objective performance. The second manuscript considered how negatively formulated items affect the quality of a survey scale while validating the English version of a promising scale for measuring website aesthetics. Here, findings suggested that the original scale was of comparable quality to versions in other languages, although an alternative version without negatively formulated items performed even better. The third manuscript examined how transparently PX researchers report on their measurement process. The manuscript showed that researchers frequently need to display more transparency when describing their measurement process, leading to recommendations addressing these issues of non-transparent reporting. Finally, the fourth manuscript reported an independent validation of a scale measuring constructs central to PX by following current best practices for psychometric scale investigation. The results generally favored the investigated scale's quality while clarifying the scale's theoretical model and identifying room for improvement regarding certain scale constructs.

Overall, the present work looked at the current state of scale-based UX research, how it should be conducted, and where there is room for improvement. In particular, findings from the four manuscripts highlighted how researchers need to pay close attention to the theories behind the measures they employ to ensure that the scales used match their own definition and conceptualization of the measured constructs. Furthermore, by demonstrating that scales validated in one context or with a particular group of participants might not perform comparably well in a new setting, the presented manuscripts have highlighted how scale validation needs to always be seen as an ongoing process. Hence, researchers should continue to re-evaluate the psychometric quality of a scale whenever possible instead of assuming that the scales used are of good psychometric quality. Because such re-evaluation is, in many cases, unrealistic (e.g., due to resource constraints), validation projects such as those reported across the presented manuscripts are crucial to scale-based UX research. Finally, the presented manuscripts further highlighted the complexity of UX research. To do justice to this complexity, UX should not be considered a singular construct that can be reduced to one number. Instead, researchers need to consider which constructs are essential to UX, both overall and within specific contexts, and how to study these constructs. For this, high-quality survey scales play a critical role in establishing a deeper understanding of what UX is and what constructs are vital to it.

References

Abbas, A., Hirschfeld, G., & Thielsch, M. T. (2022). An Arabic version of the Visual Aesthetics of Websites Inventory (AR-VisAWI): Translation and psychometric properties. *International Journal of Human–Computer Interaction*, *39*(14), 1–11. https://doi.org/10.1080/10447318.2022.2085409

Abeele, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies*, *135*, 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

Aeschbach, L. F., Opwis, K., & Brühlmann, F. (2022). Breaking immersion: A theoretical framework of alienated play to facilitate critical reflection on interactive media. *Frontiers in Virtual Reality*, *3*, 1–14. https://doi.org/10.3389/frvir.2022.846490

Aeschbach, L. F., Perrig, S. A. C., Weder, L., Opwis, K., & Brühlmann, F. (2021). Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.*, *5*(CHI PLAY). https://doi.org/10.1145/3474660

Ballou, N., Warriar, V. R., & Deterding, S. (2021). Are you open? A content analysis of transparency and openness guidelines in HCI journals. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445584

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698. https://doi.org/10.1145/1978942.1979336

Bhandari, U., Chang, K., & Neben, T. (2019). Understanding the impact of perceived visual aesthetics on user evaluations: An emotional perspective. *Information &*

Management, *56*(1), 85–93. https://doi.org/https://doi.org/10.1016/j.im.2018.07.003

Bopp, J. A., Mekler, E. D., & Opwis, K. (2016). Negative emotion, positive experience? Emotionally moving moments in digital games. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2996–3006. https://doi.org/10.1145/2858036.2858227

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. *Usability Evaluation in Industry*, *189*, 189–194.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Brühlmann, F., & Mekler, E. D. (2018). Surveys in games user research. In A. Drachen, P. Mirza-Babaei, & L. Nacke (Eds.), *Games user research* (pp. 141–162). Oxford University Press. https://doi.org/10.1093/oso/9780198794844.003.0009

Brühlmann, F., Memeti, Z., Aeschbach, L. F., Perrig, S. A. C., & Opwis, K. (2024). The effectiveness of warning statements in reducing careless responding in crowdsourced online surveys. *Behavior Research Methods*.

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022. https://doi.org/10.1016/j.metip.2020.100022

Cairns, P., & Power, C. (2018). Measuring experiences. In M. Filimowicz & V. Tzankova (Eds.), *New directions in third wave human-computer interaction: Volume 2 - methodologies* (pp. 61–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-73374-6_5

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Commun. ACM*, *63*(8), 70–79. https://doi.org/10.1145/3360311

Cockburn, A., Gutwin, C., & Dix, A. (2018). HARK no more: On the preregistration of CHI experiments. In *Proceedings of the 2018 CHI conference on human factors* *in computing systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3173574.3173715

Colusso, L., Jones, R., Munson, S. A., & Hsieh, G. (2019). A translational science model for HCI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300231

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience* (1st ed.). Harper & Row.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE publications, Inc.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, *18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology* (1st ed.). SAGE publications, Ltd.

Gault, R. H. (1907). A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, *14*(3), 366–383. https://doi.org/10.1080/08919402.1907.10532551

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Goree, S., Doosti, B., Crandall, D., & Su, N. M. (2021). Investigating the homogenization of web design: A mixed-methods approach. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445156

Graf, L., Altmeyer, M., Emmerich, K., Herrlich, M., Krekhov, A., & Spiel, K. (2022). Development and validation of a German version of the player experience inventory (PXI). *Proceedings of Mensch Und Computer 2022*, 265–275. https://doi.org/10.1145/3543758.3543763

Green, W., Dunn, G., & Hoonhout, J. (2008). Developing the scale adoption framework for evaluation (SAFE). *Proceedings of the International Workshop on Meaningful Measures: Valid Useful User Experience Measurement (VUUM).*

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction*, *19*(4), 319–349. https://doi.org/10.1207/s15327051hci1904_2

Hassenzahl, M. (2018). The thing and I (summer of '17 remix). In M. Blythe & A. Monk (Eds.), *Funology 2: From usability to enjoyment* (pp. 17–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-68213-6_2

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction*, *25*(3), 235–260. https://doi.org/10.1080/07370024.2010.500139

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, *25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Hausman, A. V., & Siekpe, J. S. (2009). The effect of web interface features on consumer online purchase intentions. *Journal Of Business Research*, *62*(1), 5–13. https://doi.org/10.1016/j.jbusres.2008.01.018

Hirsch, P. M., & Levin, D. Z. (1999). Umbrella advocates versus validity police: A life-cycle model. *Organization Science*, *10*(2), 199–212. https://doi.org/10.1287/orsc.10.2.199

Hirschfeld, G., & Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. *Ergonomics*, *58*(2), 310–320. https://doi.org/10.1080/00140139.2014.965228

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Pearson.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

Hornbæk, K., & Law, E. L.-C. (2007). Meta-analysis of correlations among usability measures. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 617–626. https://doi.org/10.1145/1240624.1240722

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

International Organization for Standardization. (2019). *ISO 9241-210:2019(en) ergonomics of human system interaction - part 210: Human-centred design for* *interactive systems* (tech. rep.). International Organization for Standardization. Vernier, Geneva, Switzerland. https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004

Juniper, E. F. (2009). Validated questionnaires should not be modified. *European Respiratory Journal*, *34*(5), 1015–1017. https://doi.org/10.1183/09031936.00110209

Kayser, D., Perrig, S. A. C., & Brühlmann, F. (2021). Measuring players' experience of need satisfaction in digital games: An analysis of the factor structure of the UPEQ. *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, 158–162. https://doi.org/10.1145/3450337.3483499

Kelley, T. L. (1927). *Interpretation of educational measurements* (1st ed.). World Book Company.

Kostakos, V. (2015). The big hole in HCI research. *Interactions*, *22*(2), 48–51. https://doi.org/10.1145/2729103

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability: Experimental analysis on the determinants of the apparent usability. *Conference Companion On Human Factors In Computing Systems*, 292–293. https://doi.org/10.1145/223355.223680

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and usability for education and work* (pp. 63–76). https://doi.org/10.1007/978-3-540-89350-9_6

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human–Computer Studies*, *60*(3), 269–298. https://doi.org/10.1016/j.ijhcs.2003.09.002

Law, E. L.-C. (2011). The measurability and predictability of user experience. *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 1–10. https://doi.org/10.1145/1996461.1996485

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 719–728. https://doi.org/10.1145/1518701.1518813

Law, E. L.-C., van Schaik, P., & Roto, V. (2014). Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies*, *72*(6), 526–541. https://doi.org/10.1016/j.ijhcs.2013.09.006

Law, E. L.-C., Brühlmann, F., & Mekler, E. D. (2018). Systematic review and validation of the Game Experience Questionnaire (GEQ) - Implications for citation and reporting practice. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 257–270. https://doi.org/10.1145/3242671.3242683

Lee, S., & Koubek, R. J. (2010). Understanding user preferences based on usability and aesthetics before and after actual use. *Interacting with Computers*, *22*(6), 530–543. https://doi.org/10.1016/j.intcom.2010.05.002

Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the system usability scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099–2102. https://doi.org/10.1145/2470654.2481287

Linxen, S., Sturm, C., Brühlmann, F., Cassau, V., Opwis, K., & Reinecke, K. (2021). How weird is CHI? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445488

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

McDonald, R. P. (1999). *Test theory: A unified treatment* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410601087

Mekler, E. D., Bopp, J. A., Tuch, A. N., & Opwis, K. (2014). A systematic review of quantitative studies on the enjoyment of digital entertainment games. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 927–936. https://doi.org/10.1145/2556288.2557078

Memeti, Z., Brühlmann, F., & Perrig, S. A. C. (2022). Lol, why do you even play? Validating the motives for online gaming questionnaire in the context of League of Legends. *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, 81–86. https://doi.org/10.1145/3505270.3558350

Miki, H. (2013). Reconsidering the notion of user experience for human-centered design. In S. Yamamoto (Ed.), *Human interface and the management of information. information and interaction design* (pp. 329–337). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39209-2_38

Minge, M., & Thüring, M. (2018). Hedonic and pragmatic halo effects at early stages of user experience. *International Journal of Human-Computer Studies*, *109*, 13–25. https://doi.org/10.1016/j.ijhcs.2017.07.007

Moosbrugger, H., & Kelava, A. (2000). *Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]* (3rd ed.). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-61532-4

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal Of Human-Computer Studies*, *68*(10), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

Nacke, L., & Drachen, A. (2011). Towards a framework of player experience research. *Proceedings of the second international workshop on evaluating player experience in games at FDG*, *11.*

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). Mcgraw hill book company.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oulasvirta, A., & Hornbæk, K. (2016). HCI research as problem-solving. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4956–4967. https://doi.org/10.1145/2858036.2858283

Pengnate, S., Sarathy, R., & Lee, J. (2019). The engagement of website initial aesthetic impressions: An experimental investigation. *International Journal of Human–Computer Interaction*, *35*(16), 1517–1531. https://doi.org/10.1080/10447318.2018.1554319

Perrig, S. A. C., Aeschbach, L. F., Scharowski, N., von Felten, N., Opwis, K., & Brühlmann, F. (2022). Measurement practices in UX research: A systematic quantitative literature review. https://doi.org/10.31234/osf.io/3jz67

Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544549.3585808

Perrig, S. A. C., Scharowski, N., Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2023). Independent validation of the player experience inventory: Findings from a large set of video game players. *Manuscript submitted for publication.*

Perrig, S. A. C., Ueffing, D., Opwis, K., & Brühlmann, F. (2023). Smartphone app aesthetics influence users' experience and performance. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1113842

Perrig, S. A. C., von Felten, N., Honda, M., Opwis, K., & Brühlmann, F. (2023). Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-pos. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2023.2258634

Pettersson, I., Lachner, F., Frison, A.-K., Riener, A., & Butz, A. (2018). A bermuda triangle? A review of method application and triangulation in user experience evaluation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3173574.3174035

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

Rogerson, M. J., Gibbs, M. R., & Smith, W. (2018). Cooperating to compete: The mutuality of cooperation and competition in boardgame play. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173767

Roto, V. (2009). Demarcating user experience. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. O. Prates, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2009* (pp. 922–923). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03658-3_112

Ryan, R. M., & Deci, E. L. (2000). *Intrinsic motivation inventory (IMI)*. Center for Self-Determination Theory. https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of personality and Social Psychology*, *45*(4), 736–750. https://doi.org/10.1037/0022-3514.45.4.736

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*, 344–360. https://doi.org/10.1007/s11031-006-9051-8

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–199. https://doi.org/10.7334/psicothema2014.266

Saremi, M., Sadeghi, V., Khodakarim, S., & Maleki-Ghahfarokhi, A. (2023). Farsi version of Visual Aesthetics of Website Inventory (FV-VisAWI): Translation and psychometric evaluation. *International Journal of Human–Computer Interaction*, *39*(4), 834–841. https://doi.org/10.1080/10447318.2022.2049138

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1609–1618. https://doi.org/10.1145/1518701.1518947

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2215–2224. https://doi.org/10.1145/1978942.1979266

Scharowski, N., & Perrig, S. A. C. (2023). Distrust in (X)AI – Measurement artifact or distinct construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams*. https://doi.org/10.48550/arXiv.2303.16495

Scharowski, N., Perrig, S. A. C., Aeschbach, L. F., von Felten, N., Opwis, K., & Wintersberger, F., Philippand Brühlmann. (2023). To trust or distrust trust measures: Validating questionnaires for trust in AI. *Manuscript submitted for publication*.

Schrepp, M., Otten, R., Blum, K., & Thomaschewski, J. (2021). What causes the dependency between perceived aesthetics and perceived usability? *International Journal of Interactive Multimedia & Artificial Intelligence*, *6*(6), 78–85. https://doi.org/10.9781/ijimai.2020.12.005

Seckler, M., Heinz, S., Forde, S., Tuch, A. N., & Opwis, K. (2015). Trust and distrust on the web: User experiences and website characteristics. *Computers in Human Behavior*, *45*, 39–50. https://doi.org/10.1016/j.chb.2014.11.064

Seng, T. L., & Mahmoud, M. A. S. (2020). Perceived e-service quality and e-store loyalty: The moderated mediating effect of webpage aesthetics and e-customer satisfaction. *International Journal of Advanced and Applied Sciences*, *7*(5), 111–117. https://doi.org/10.21833/ijaas.2020.05.014

Skulmowski, A., Augustin, Y., Pradel, S., Nebel, S., Schneider, S., & Rey, G. D. (2016). The negative impact of saturation on website trustworthiness and appeal: A temporal model of aesthetic website perception. *Computers in Human Behavior*, *61*, 386–393. https://doi.org/https://doi.org/10.1016/j.chb.2016.03.054

Sonderegger, A., & Sauer, J. (2010). The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied Ergonomics*, *41*(3), 403–410. https://doi.org/10.1016/j.apergo.2009.09.002

Tenzer, F. (2023). *Anzahl der Smartphone-Nutzer weltweit von 2016 bis 2022 und Prognose bis 2028 [Number of smartphone users worldwide from 2016 to 2022 and forecast to 2028]*. Statista. https://de.statista.com/statistik/daten/studie/309656/umfrage/prognose-zur-anzahl-der-smartphone-nutzer-weltweit/

Thielsch, M. T., & Hirschfeld, G. (2021). Quick assessment of web content perceptions. *International Journal of Human–Computer Interaction*, *37*(1), 68–80. https://doi.org/10.1080/10447318.2020.1805877

Thielsch, M. T., Scharfen, J., Masoudi, E., & Reuter, M. (2019). Visual aesthetics and performance: A first meta-analysis. In *Proceedings of Mensch Und Computer 2019* (pp. 199–210). Association for Computing Machinery. https://doi.org/10.1145/3340764.3340794

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Tractinsky, N. (1997). Aesthetics and apparent usability: Empirically assessing cultural and methodological issues. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 115–122. https://doi.org/10.1145/258549.258626

Tractinsky, N. (2018). The usability construct: A dead end? *Human–Computer Interaction*, *33*(2), 131–177. https://doi.org/10.1080/07370024.2017.1298038

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. https://doi.org/10.1016/S0953-5438(00)00031-X

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596–1607. https://doi.org/https://doi.org/10.1016/j.chb.2012.03.024

Vermeeren, A. P. O. S., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: Current state and development needs. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521–530. https://doi.org/10.1145/1868914.1868973

von Felten, N., Brühlmann, F., & Perrig, S. A. C. (2022). Independent validation of the Video Game Dispositional Flow Scale with League of Legends players. *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, 44–50. https://doi.org/10.1145/3505270.3558351

Wacharamanotham, C., Eisenring, L., Haroz, S., & Echtler, F. (2020). Transparency of CHI research artifacts: Results of a self-reported survey. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376448

Wiemeyer, J., Nacke, L., Moser, C., & 'Floyd' Mueller, F. (2016). Player experience. In R. Dörner, S. Göbel, W. Effelsberg, & J. Wiemeyer (Eds.), *Serious games: Foundations, concepts and practice* (pp. 243–271). Springer International Publishing. https://doi.org/10.1007/978-3-319-40612-1_9

Acknowledgments

I am sincerely thankful to numerous people for supporting me on my way to this thesis. Without their help, this work would not have been possible:

- Klaus Opwis, my PhD supervisor, for the opportunity to write this thesis and for his continuous guidance, support, trust, and encouragement during these years.
- Javier Andrés Bargas-Avila, for volunteering to be my second supervisor and for introducing me to HCI during my Bachelor's studies.
- Florian Brühlmann, the former head of the Human-Computer Interaction Research Group, for his guidance, support, countless conversations, and for sparking my interest in scale development, validation, and theory.
- My co-authors, Lena Fanya Aeschbach, Nicolas Scharowski, Nick von Felten, David Ueffing, Zgjim Memeti, Marimo Honda, Lorena Weder, and Philipp Wintersberger.
- My colleagues at the Center for General Psychology and Methodology, Antony Berbert de Castro Hüsler, Ariane Haller, Beat Vollenwyder, Léane Wettstein, and Melanie Svab.
- And most of all, I would like to thank my partner, Jeanne Wenger, along with my family, Andrea Heeb Perrig and Thierry Perrig, who supported me during my PhD throughout easy as well as hard times.

curriculum vitae

Sebastian A. C. Perrig

Aus Datenschutzgründen entfernt.

Appendix

1. **Perrig, S. A. C.**, Ueffing, D., Opwis, K., & Brühlmann, F. (2023). Smartphone app aesthetics influence users' experience and performance. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1113842

2. **Perrig, S. A. C.**, von Felten, N., Honda, M., Opwis, K., & Brühlmann, F. (2023). Development and validation of a positive-item version of the Visual Aesthetics of Websites Inventory: The VisAWI-pos. *International Journal of Human–Computer Interaction.* https://doi.org/10.1080/10447318.2023.2258634

3. Aeschbach, L. F., **Perrig, S. A. C.**, Weder, L., Opwis, K., & Brühlmann, F. (2021). Transparency in measurement reporting: A systematic literature review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY). https://doi.org/10.1145/3474660

4. **Perrig, S. A. C.**, Scharowski, N., Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2023). Independent validation of the Player Experience Inventory: Findings from a large set of video game players. *Manuscript submitted for publication.*

TYPE Original Research
PUBLISHED 14 June 2023
DOI 10.3389/fpsyg.2023.1113842

OPEN ACCESS

EDITED BY
Carlos Duarte,
University of Lisbon, Portugal

REVIEWED BY
Wayne Giang,
University of Florida, United States
Laura Belli,
University of Parma, Italy

*CORRESPONDENCE
Sebastian A. C. Perrig
✉ sebastian.perrig@unibas.ch

†These authors have contributed equally to this work and share first authorship

RECEIVED 01 December 2022
ACCEPTED 19 May 2023
PUBLISHED 14 June 2023

CITATION
Perrig SAC, Ueffing D, Opwis K and Brühlmann F (2023) Smartphone app aesthetics influence users' experience and performance. *Front. Psychol.* 14:1113842.
doi: 10.3389/fpsyg.2023.1113842

Smartphone app aesthetics influence users' experience and performance

Sebastian A. C. Perrig*†, David Ueffing†, Klaus Opwis and Florian Brühlmann

Human-Computer Interaction Research Group, Center for General Psychology and Methodology, Faculty of Psychology, University of Basel, Basel, Switzerland

Past research has demonstrated that aesthetics affect users' experiences in various ways. However, there is little research on the impact of interface aesthetics on user performance in a smartphone app context. The present paper addresses this research gap using an online experiment ($N = 281$). Two variants of the same web app were created and manipulated in their aesthetics. Participants were randomly assigned to either variant and asked to explore the app before answering questions concerning the app's content. Results showed a significant positive effect of aesthetics on perceived usability and aesthetics. Furthermore, results point toward a positive impact of interface aesthetics on performance (i.e., the number of questions answered correctly). Thus, results indicate that a visually appealing smartphone web app increases users' subjective experience and objective performance compared to an unaesthetic app. This suggests that user interface aesthetics impact users' experiences and provide stakeholders with quantifiable value and competitive advantage.

KEYWORDS

aesthetics, performance, usability, mobile devices, smartphones, User Experience (UX)

1. Introduction

Smartphone use is developing rapidly worldwide. While there were 2.49 billion active smartphone users in 2016, this number has risen to 3.6 billion in the following 4 years, and by 2024, 4.5 billion active users are expected (Tenzer, 2022). Furthermore, 54.97% of all website visits worldwide in 2021 were made via smartphones (Statista Research Department, 2022) and smartphones are expected to replace computers in certain areas of daily life (Anderson, 2019). It is, therefore, not surprising that many software developers frequently develop mobile device applications (apps) or port their computer programs to them. A shift in focus by developers and businesses from computer programs to apps has resulted in the ability to perform almost any daily task with an app, ranging from contacting friends to banking transactions. There appears to be an app for each activity, or a whole market of specific apps for each task, resulting in a competitive market where users can choose between various alternatives. Given the omnipresence of smartphone apps in private and professional life, the question arises as to what makes a smartphone app successful in such a highly competitive market. Several indications point to aesthetics, which has a multi-layered influence on people's perceptions. An example of this is the influence of the aesthetics of an app on users' subjective evaluation, which can take place within fractions of a second (Guo et al., 2020). It is thus unsurprising that in the developer community and human-computer interaction (HCI) field, more and more attention is being paid to aesthetics (Tractinsky and Hassenzahl, 2005). Several studies have shown a positive effect of aesthetics on subjective perception and the resulting reactions (De Angeli et al., 2006; Thüring and Mahlke, 2007; Douneva et al., 2016). Furthermore, some researchers have already demonstrated that aesthetics positively affects performance in various contexts (Salimun et al., 2010; Sonderegger and Sauer, 2010; Reppa and McDougall, 2015). However, to our knowledge, there is still limited empirical investigation into the effects of aesthetics within a smartphone app context, despite the growing importance of the mobile device market. It thus remains unclear to what extent past findings concerning the impact of aesthetics on the users' experiences and performance can also be found within the smartphone device context. The present study thus investigated the effect of smartphone app aesthetics on users' subjective perception of aesthetics and usability and users' performance with an experimental study to address this research gap.

2. Related work

2.1. A brief excursion into the world of apps

Mobile interfaces differ substantially from desktop websites (Nielsen and Budiu, 2013). For example, given the smaller screen size, less information can be displayed simultaneously, and while exact clicking on smaller targets is possible with precise mouse movements on desktop websites, less precision is possible on smaller smartphone touch screens (Nielsen and Budiu, 2013). Thus, while we might assume that results from a desktop setting also apply to a smartphone context, we can only be sure once an empirical investigation is conducted. In addition, past research has already shown that non-smartphone mobile devices differ from desktop websites concerning the effect of aesthetics on performance (Thielsch et al., 2019b). Smartphones, however, which differ from past mobile devices (e.g., because of touchscreens), have not yet been studied in this respect. Further, Groth and Haslwanter (2015) found significant differences in perceived usability and user experience between desktop computers and smartphones, while Nielsen and Budiu (2013) found lower e-commerce conversion rates for mobile phones in contrast to desktop computers, and Zhu et al. (2020) showed that written user reviews differ between mobile and desktop devices in several aspects (e.g., fewer words and more pictures). Thus, past research has shown that results from a desktop setting can differ from those found in a mobile context, but the effect of aesthetics on performance still needs to be determined for smartphone apps.

Although the term *app* is used frequently, it does not always imply the same thing. According to the Merriam-Webster Dictionary, the term *application* refers to "a program (such as a word processor or a spreadsheet) that performs a particular task or set of tasks."[1] In contrast, the term *app* describes "an application designed for a mobile device (such as a smartphone)."[2] A further distinction is made between native and web apps (Jobe, 2013). A native app is downloaded from a store and permanently installed on the smartphone, with a separate app programmed for each platform (El-Kassas et al., 2017). On the other hand, a web app is a particular form of an interactive website that behaves like a conventional application but does not have to be installed on a smartphone, which is a great advantage of web apps (Jobe, 2013). In the case of mobile versions of a website, the term *generic mobile web application* refers to versions of a website either developed for a mobile context or adapted through responsive design (Jobe, 2013). Web apps can be used across platforms and do not require custom programming for each operating system. In addition, developers can distribute updates to all users faster and more efficiently, as there is no need to trigger a manual update process as with native apps (Liu et al., 2015). Studies have also shown that web apps perform better than native apps under certain conditions (Jobe, 2013; Liu et al., 2015; Ma et al., 2017). Large companies increasingly recognize these advantages of web apps over native apps to better reach and support users. While Google is moving forward with plans to foster web apps,[3] Microsoft released its game streaming platform *Xbox Cloud Gaming* as a web app for multiple platforms.[4] Similarly, Apple allows developers to launch applications as web apps (Apple Pty Ltd., 2021). Experts, therefore, agree that web apps will increasingly be found on the market in the future, offering an excellent alternative to native apps (Ater, 2017).

2.2. Aesthetics in HCI

Initiated by works such as Kurosu and Kashimura (1995) or Tractinsky et al. (2000), aesthetics has been extensively investigated within the field of HCI. Past research provided evidence that visually appealing websites are perceived as more trustworthy (Lindgaard et al., 2011) and that user purchase intent increases with more appealing systems (Hausman and Siekpe, 2009), as do satisfaction (Lindgaard, 2007) and preference (Lee and Koubek, 2010). From a psychological point of view, aesthetics appear to satisfy basic human needs of enjoyment and wellbeing (Postrel, 2004). Furthermore, when it comes to self-expression, users can express their individuality by personalizing interfaces or lock screens, allowing them to differentiate themselves from others (Hassenzahl, 2018). Lee and Koubek (2010) further showed that users initially evaluate an interactive system significantly based on its aesthetic impression, while Wiecek et al. (2019) found that product aesthetics (e.g., smartphone cases) had a positive effect on usage intensity while deterring users from switching to different products. Over the past two decades, such promising research results have enabled designers and the HCI community to move away from initial concerns by some (e.g., Andre and Wickens, 1995) that aesthetic design interferes with work objectives. Aesthetics is now a widely recognized "must-have" factor that gets a great deal of attention when developing systems (Thielsch et al., 2014).

2.2.1. Perceived visual aesthetics

Moshagen and Thielsch (2010) defined aesthetics "as an immediate pleasurable subjective experience that is directed toward an object and not mediated by intervening reasoning"

[1] https://www.merriam-webster.com/dictionary/app

[2] https://www.merriam-webster.com/dictionary/application

[3] https://www.youtube.com/watch?v=GSiUzuB-PoI

[4] https://www.xbox.com/en-US/xbox-game-pass/cloud-gaming

(p. 690). According to Lavie and Tractinsky (2004), aesthetics can be separated into classic and expressive aesthetics. Classic aesthetics refers to clean, pleasant, and symmetrical attributes, while expressive aesthetics refers to characteristics such as creative, original, and sophisticated. Moshagen and Thielsch (2010) further argued that the construct of visual aesthetics is represented by four facets: simplicity, diversity, colorfulness, and craftsmanship. Simplicity describes concepts like unity or homogeneity, while diversity represents aspects such as novelty and creativity. Simplicity correlates highly with classic, and diversity correlates highly with expressive aesthetics of Lavie and Tractinsky (2004). Colorfulness considers aspects such as the placement and combination of colors. Finally, craftsmanship reflects whether the product has a harmonious design and uses modern technologies. Given that multiple studies have investigated this conceptualization of aesthetics (e.g., Moshagen and Thielsch, 2010, 2013) where it has proven itself useful, this paper will follow this definition by Moshagen and Thielsch (2010).

2.2.2. Objective facets of aesthetics

Examining aesthetics raises the question of how products can be objectively manipulated to realize different aesthetic impressions. Various studies have shown two salient characteristics, complexity and symmetry, to strongly influence the perception of websites (Bauerly and Liu, 2008; Lai et al., 2010; Tuch et al., 2010; Bi et al., 2011; Seckler et al., 2015). Moreover, they proved to be some of the most distinctive design features upon initial observation (Leder et al., 2004). Bauerly and Liu (2008) postulated that symmetry helps viewers structure content by creating regular and meaningful forms. Moreover, in Seckler et al. (2015), symmetry was the biggest influencing factor on the subjective overall aesthetic perception. In contrast, complexity is more challenging to define (Xing and Manning, 2005). Nevertheless, several studies described visual complexity by the quantity of objects, clutter, openness, symmetry, organization, and variety of colors (Olivia et al., 2004; Michailidou et al., 2008; Riegler and Holzmann, 2018). Based on this definition, multiple HCI studies provided evidence for a negative linear correlation between visual complexity and aesthetic perception, implying that higher complexity leads to lower aesthetic ratings (Michailidou et al., 2008; Tuch et al., 2012a; Seckler et al., 2015).

Besides complexity and symmetry, color was repeatedly shown to be among the most striking design features at first glance (Cyr et al., 2010; Reinecke et al., 2013). In the context of HCI, color is frequently represented by the Hue-Saturation-Brightness (HSB) model, according to which color is composed of three parts: hue, saturation, and brightness (Smith, 1978). Hue is defined as a pure, spectral color such as blue, red, or yellow. In various studies, blue and gray websites were rated as the most attractive and yellow and purple as the least attractive ones (Cyr et al., 2010; Seckler et al., 2015). Comparable results have also been found in studies not related to HCI (Fortmann-Roe, 2013; Palmer et al., 2013; Oyibo and Vassileva, 2020). Saturation, the second aspect of the HSB model, describes the intensity of the color, which has not been extensively researched to date (Seckler et al., 2015). Nevertheless, there is an indication that western adults generally prefer higher saturated websites (Palmer and Schloss, 2010; Lindgaard et al., 2011; Seckler et al., 2015). Brightness, the last aspect, describes the perceived luminance of a color. As with saturation, there is little scientific evidence on the effects of brightness (Seckler et al., 2015). However, some evidence indicates that websites with high background luminance are rated as the most beautiful (Palmer and Schloss, 2010; Lindgaard et al., 2011).

2.2.3. Effects of aesthetics on usability

The positive effect of aesthetics on various subjective aspects of users' experiences, such as preferences and trust (Moshagen and Thielsch, 2010), user satisfaction (Tractinsky et al., 2000; Lavie and Tractinsky, 2004; Tseng and Lee, 2019), or joy of use (Lingelbach et al., 2022) has already been demonstrated and widely researched. Another frequently studied subject is the effect of aesthetics on usability. The International Organization for Standardization (2018) defines *system usability* as "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." In research, a distinction is made between subjective and objective usability. Subjective usability concerns users perception and attitudes regarding a system, while measures of objective usability evaluate a systems properties not dependent on a persons perception (Hornbæk, 2006). Researchers, therefore, addressed the question of what subjectively perceived usability depends on. Several studies have found a robust effect of aesthetics on subjective usability, showing that users working with a more attractive system rated it as more usable than users of a less attractive one (Moshagen et al., 2009; Sonderegger and Sauer, 2010; Sonderegger et al., 2014; Gu et al., 2016; Minge and Thüring, 2018; Otten et al., 2020; Schrepp et al., 2021).

2.3. Aesthetics and performance—current state of research

Prompted by aesthetics' effects on users' subjective experiences, the question of whether visual aesthetics also influence an objective construct such as performance arose. In this paper, *performance* is defined in line with Thielsch et al. (2019b) as "an objectively measurable outcome of a user's interplay with a website, software or other interactive system" (p. 200). While there is initial evidence for an effect of aesthetics on performance, it is not yet clear whether users only believe that they perform better with a more aesthetic application or whether there is an objectively measurable change in performance. Research results thus far are ambivalent (Thielsch et al., 2019b). Some studies support a performance improvement when interacting with an aesthetically more appealing interface (Sonderegger and Sauer, 2010; Douneva et al., 2016; Baughan et al., 2020; Reppa et al., 2021), whereas others show a contrary effect (Sauer and Sonderegger, 2011; Sonderegger et al., 2014). In addition, several studies could not show any significant effect (Douneva et al., 2015; Gu et al., 2016; Thielsch et al., 2019a). Given these contradictory findings, various explanations have been made to understand aesthetics' effect on performance, summarized in the following section.

2.3.1. Theoretical considerations

Szabo and Kanuka (1999) postulated that good design improves performance by reducing cognitive processing effort. This reduced effort is achieved because good design enables faster recognition of visual objects. In this regard, good design is implemented through low complexity and higher coherence, promoting the automatic processing of information. Bad design, on the other hand, provokes more inefficient, manual processing (Szabo and Kanuka, 1999). Inspired by this idea, various researchers have discussed attentional effects of aesthetic design (e.g., Reppa et al., 2008). In this context, additional cognitive effects of website perception have been debated, such as visual complexity and prototypicality, bottom-up perception processes, and mental models (Tuch et al., 2009; Douneva et al., 2016).

Tractinsky et al. (2000) took the idea of the halo effect from Psychology[5] and postulated that "what is beautiful is usable," arguing that the user infers from the aesthetic design to other parts of the application. For example, due to the halo effect, the user initially perceives an application as aesthetic and concludes from this judgment alone that the application has good functionality. Some studies provided evidence for this assumption (Lavie and Tractinsky, 2004; Hartmann et al., 2008; Quinn and Tran, 2010), while others found a reversed effect under certain conditions (Tuch et al., 2012b).

Sonderegger and Sauer (2010) argued that aesthetic design puts users at ease or in a kind of "flow state" (Csikszentmihalhi, 1997). In this state, users perceive the tasks given to them as congruent with their abilities, leading to faster processing and increased motivation when using a system, consequently increasing performance. This is especially the case in a work context. They further claimed that users focus on a design that is subjectively perceived as beautiful and then "lose themselves" in it, leading to more inefficient processing and, thus, lower performance. Users in such situations are no longer fully focused on the task but try to prolong the pleasant experience of interacting with the appealing design. This "prolongation of joyful experience" occurs more often in leisure tasks, focusing on fun and enjoyment rather than performance (Sonderegger and Sauer, 2010; Sonderegger et al., 2014).

Overall, there are few systematic studies on these explanatory concepts (Thielsch et al., 2019b), and results on the relationship between aesthetics and performance are often contradictory. Thielsch et al. (2019b) have taken this as an occasion to conduct a meta-analysis. Results revealed a small, positive effect of interface aesthetics on user performance ($g = 0.12$). Moreover, a complementary finding was that more aesthetically pleasing variants significantly impact user performance, especially when interacting with mobile devices and software applications. However, the studies and data available to date are far from adequate, leading the authors to formulate a call to action for more substantiated research.

2.4. Study goals

As Thielsch et al. (2019b) suggested in their meta-analysis, aesthetics influence user performance in the context of digital products. However, their results should be regarded with caution, as there were several challenges with the included studies. First, the authors emphasized that there are still too few high-quality publications that address the relationship between aesthetics and performance. Therefore, further research is essential to understand aesthetics' effect on user performance better. Furthermore, previous studies have primarily focused on computer applications. However, smartphones, with their smaller displays and on-the-go use, have unique requirements and strengths (Adepu and Adler, 2016). Thus, previous findings on computer interfaces may not directly apply to smartphone interfaces and apps. Research addressing mobile devices to date mainly focused on the external appearance of the device as an aesthetic manipulation (e.g., Sonderegger and Sauer, 2010; Sonderegger et al., 2014; Minge and Thüring, 2018). Thus, there is a lack of studies centering on mobile devices' interfaces.

The present work addresses these issues by focusing solely on an app's user interface rather than a smartphone's exterior design. The specific device used by participants was not considered as long as participants used a smartphone device to access the online study. Specifically, this study examined the impact of an app's interface aesthetics on user performance during use. To investigate aesthetics, we employed the definition of Moshagen and Thielsch (2010, 2013). Perceived usability and aesthetics were measured using two validated survey scales. A set of self-developed knowledge questions related to the app's content filled out post-interaction were used to quantify performance. Overall, this study aimed to address the current research gap by investigating the effect of interface aesthetics on performance in the context of mobile devices. The results promote a deeper understanding of user performance and behavior in the context of smartphone use and the influence of aesthetics on such interactions.

2.4.1. Research hypotheses

We derived the following three research hypotheses based on the study goals and previous research described above:

- H1: Concerning perceived usability, users of the aesthetically pleasing variant of the app will exhibit higher levels of subjective usability than users of the unaesthetic one.
- H2: Concerning task completion time, users of the aesthetic variant of the app will complete tasks related to the app content faster than users of the unaesthetic variant.
- H3: Considering task performance, reflected in a performance score, users interacting with the aesthetic variant of an app will have a higher performance score, compared to those interacting with the unaesthetic variant.

3. Materials and methods

To achieve our research goals, we conducted a between-subjects design online experiment. Participants interacted with one of two variants of a fictitious event agency's web app. The two variants

[5] In Psychology, the halo effect refers to a phenomenon where certain characteristics, such as physical beauty, are perceived early in an interaction, consecutively influencing the perception of other personal characteristics (Thorndike, 1920; Dion et al., 1972).

of the app were manipulated in terms of aesthetics to investigate a possible relationship between the app's aesthetics and the user's performance and experience during the interaction.

3.1. Sample

We recruited an initial sample of 387 participants over Amazon Mechanical Turk (MTurk),[6] out of which 344 completed the online experiment. Ethical review and approval was not required for the study in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Only workers located in the United States of America with a human-intelligence-task approval of 95% and at least 100 approved tasks were allowed to participate in the experiment. For data cleaning purposes, we imposed several criteria on the sample. First, all subjects who indicated a visual or color impairment were removed ($n = 22$) because participants had to perceive and evaluate aesthetics manipulated by color, among other things. Following recommendations by Brühlmann et al. (2020), we removed one participant for failing to correctly answer an attention check item (Meade and Craig, 2012; Curran, 2016), and one respondent because they self-reported that their data should not be used due to insufficient quality (Meade and Craig, 2012). Seven participants were removed due to interruptions while answering the survey. Furthermore, we removed five participants for responding to the Visual Aesthetics of Websites Inventory (VisAWI, Moshagen and Thielsch, 2010) and Usability Metric for User Experience (UMUX, Finstad, 2010) too quickly (following Huang et al., 2012) and 20 participants who took too long to answer the survey (outliers concerning response time based on the interquartile range). Finally, we removed seven participants with a suspicious amount of the same answers for the VisAWI and UMUX, indicating that they ignored the reverse-coded answers (i.e., same answers not only across all positively formulated items but also for reversed items). After data cleaning, a final sample of 281 complete responses remained (aesthetic condition = 139, unaesthetic condition = 142). Participants self-reported an average age of 35.39 years [standard deviation (SD) = 9.77, *range* = 18–70] and 137 participants identified as female (male = 135, non-binary = 5, preferred not to answer = 4).

3.2. Materials and experimental manipulations

To reveal possible effects of aesthetics on performance, following the findings of Thielsch et al. (2019b), different variants of the same app were created and manipulated to be either as aesthetically pleasing or as unaesthetic as possible. In line with past research, we opted for manipulating aesthetics as much as possible to avoid problems caused by weak manipulation (Thielsch et al., 2019a). For the final study, two variants of the same app (Figure 1) were developed using the free website development platform Wix.[7] Care was taken to keep all aspects of the app not related directly to aesthetics the same, including avoiding strong manipulations of system usability. Therefore, we purposefully refrained from altering system properties related to usability in past research, such as manipulation of the information architecture (e.g., menu labels as in Tuch et al., 2012b), menu structure (as in Minge and Thüring, 2018) or page response time (e.g., system delay as in Tractinsky et al., 2000). Aesthetics was thus manipulated in line with past research by manipulating symmetry and color combinations (e.g., Minge and Thüring, 2018) or changing the website structure, color, and fonts while keeping the content constant (as in Iten et al., 2018). In addition, we considered the Web Content Accessibility Guidelines (Accessibility Guidelines Working Group, 2018) to keep both variants as comparable as possible. For example, the contrast ratios of the elements for both variants were always at least level AA according to the guidelines. In general, the base variant of the app before manipulation was designed to be as realistic as possible. In addition, efforts were made to maximize the difference in aesthetics between the two final variants of the app. The following subsections describe the development of the two app variants in more detail.

3.2.1. Initial stimuli design

Feedback was gathered from a team of experts during various stages of the design process to ensure a realistic app design. Specifically, four user interface and user experience designers were consulted, and their feedback was incorporated into the development of the apps. These experts contributed their expertise in aesthetic and user-centered software design in individual discussions. This way, efforts were made to develop a realistic and well-executed initial app. This base app was then manipulated regarding aesthetics, based on the conceptualization of aesthetics by Moshagen and Thielsch (2010), to create seven different app variants. For creating these app variants, three aspects of aesthetics were varied: color, complexity, and symmetry. Different color combinations were used, shown to be perceived by users as particularly aesthetic or unaesthetic in past research (Seckler et al., 2015). Different amounts of colors were included in the color scheme of the respective app variant to manipulate complexity. Furthermore, the number of fonts was varied to alter the consistency of the app variants, and thus the complexity of the overall appearance (Thielsch et al., 2019a). Symmetry was manipulated mainly by deviating from the central vertical axis of the screen.

3.2.2. Preliminary stimuli evaluation

The seven initial app variants were compared in a preliminary evaluation to select the variants with the highest and lowest aesthetics ratings as stimuli for the main study. A total of 12 HCI researchers (master's and Ph.D. students enrolled in the HCI program at the authors' university) rated screenshots for each of the seven app variants using the four-item short version of the VisAWI, the VisAWI-S (German version, Moshagen and Thielsch, 2013).[8] In addition, participants answered an ordering question that asked

[6] https://www.mturk.com

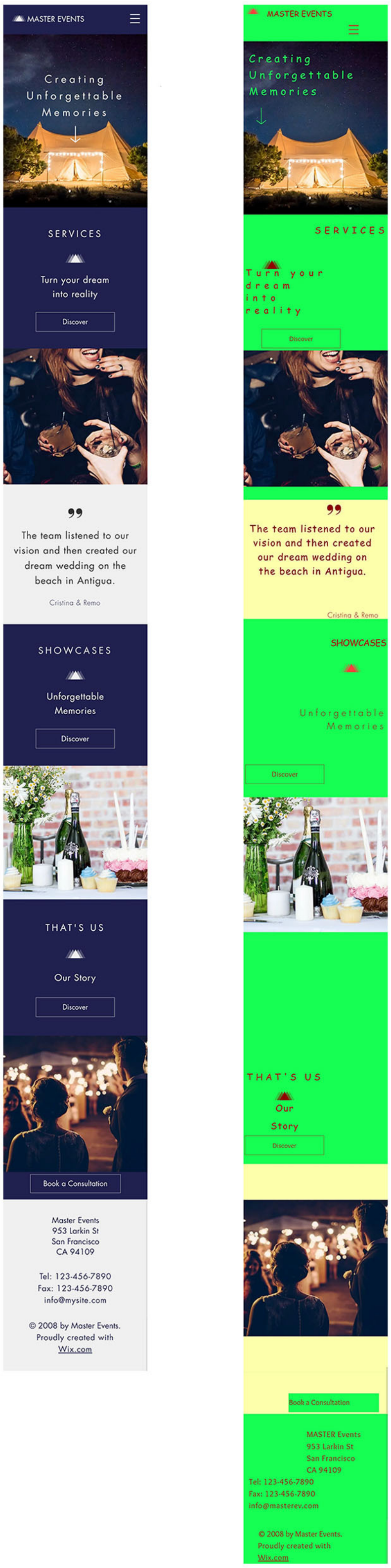
[7] https://wix.com

[8] The German version of the VisAWI-S was used in the preliminary investigation because the participants were German-speaking. However, the

for all variants to be sorted from highest to lowest aesthetics. The VisAWI-S score of the app variant rated highest [mean (*M*) = 5.23, *SD* = 1.23] exceeded the cut-off of 4.5 for an aesthetic design by Hirschfeld and Thielsch (2015) and differed clearly from the variant rated lowest (*M* = 2.25, *SD* = 1.02). Ratings from the ordering question were also consistent with the VisAWI-S ratings. Furthermore, we performed a one-way analysis of variance (ANOVA) to compare the app variants' effect on the VisAWI-S score. Results revealed a statistically significant difference between at least two variants [$F_{(6,\,77)} = 14.10$, $p < 0.0001$, $\eta^2 = 0.52$]. Because the VisAWI-S score was not normally distributed, we further calculated a Kruskal-Wallis rank sum test, which also showed a significant difference [$\chi^2(6) = 44.07$, $p < 0.0001$]. Finally, Tukey's Honest Significant Difference Test for multiple comparisons showed that the mean value was significantly different between the app variant rated highest and the variant rated lowest [$p < 0.0001$, difference in means = 2.98, 95% CI $(1.59, 4.37)$].

3.2.3. Final stimuli used

Figure 1 shows the two final app variants used in the main experiment. For the aesthetic variant, based on findings by Seckler et al. (2015), only the colors blue and gray were used (see Supplementary material for exact color codes).[9] In addition, we used only one font type (*Futura*) across the app. Due to the small number of colors and only one font, we considered this condition of low complexity. We kept symmetry at a maximum throughout the app. Every element was aligned around a vertical, central axis, and care was taken to ensure that each element occupied approximately the same amount of space. In the unaesthetic variant, six different color variations were chosen based on Seckler et al. (2015), including three shades of red. Furthermore, we used three different fonts across the app (*Comic Sans MS, Overlock*, and *Futura Light*). Thus, the complexity in this app variant was arguably higher than in the aesthetic variant. Wherever possible, symmetry was purposefully disregarded. Emphasis was placed on arranging the various surface objects as asymmetrically as possible so that no symmetry or pattern could be discerned.



MASTER EVENTS

Creating Unforgettable Memories

SERVICES

Turn your dream into reality

Discover

"The team listened to our vision and then created our dream wedding on the beach in Antigua."

Cristina & Remo

SHOWCASES

Unforgettable Memories

Discover

THAT'S US

Our Story

Discover

Book a Consultation

Master Events
953 Larkin St
San Francisco
CA 94109

Tel: 123-456-7890
Fax: 123-456-7890
info@mysite.com

© 2008 by Master Events.
Proudly created with
Wix.com

MASTER EVENTS

Creating Unforgettable Memories

SERVICES

Turn your dream into reality

Discover

"The team listened to our vision and then created our dream wedding on the beach in Antigua."

Cristina & Remo

SHOWCASES

Unforgettable Memories

Discover

THAT'S US

Our Story

Discover

Book a Consultation

MASTER Events
953 Larkin St
San Francisco
CA 94109
Tel: 123-456-7890
Fax: 123-456-7890
info@masterev.com

© 2008 by Master Events.
Proudly created with
Wix.com

FIGURE 1
The two final variants of the web app used as stimuli in this study. Shown is the landing page of the aesthetic **(left)** and the unaesthetic **(right)** implementation. Images used from Unsplash. Note that the first image depicted in the screenshots was replaced with a comparable image for this publication due to copyright.

3.3. Measurements

Two validated self-reported survey scales from previous research were used for data collection alongside two indicators of performance (performance score, performance time). Before interpreting the data, we investigated the scales' reliability and validity to ensure the quality of our measurements, which should always be done whenever scales are used with a new sample (Furr, 2011). The scale used to measure aesthetics was not previously validated in its English version but only in German with German-speaking participants (Abbas et al., 2022). The scale's quality in

app screenshots rated were in English because they were designed to be used with English-speaking participants in the main study.

[9] https://osf.io/xsdqy

English was thus unclear. In addition, both scales were developed with non-mobile devices, so we wanted to ensure sufficient scale quality in our context before interpreting the results. Reliability was investigated using two measures of internal consistency, coefficients α (Cronbach, 1951) and ω (McDonald, 1999). Regarding validity, we investigated the structure of all survey scales using confirmatory and exploratory factor analysis. The essential parts of these investigations are reported as part of the following subsections, while full details are provided on the Open Science Framework (OSF).[10]

3.3.1. Perceived visual aesthetics: the VisAWI

The VisAWI (Moshagen and Thielsch, 2010) was used to measure the perceived visual aesthetics of the app. The VisAWI is a self-reported survey scale comprising 18 items (including eight negatively formulated items) distributed over four subscales: *Simplicity, diversity, colorfulness*, and *craftsmanship*. Ratings were made on a 7-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). Scale values for the subscales were formed by calculating means across items for each subscale, while the overall score was calculated by adding up the four subscale values and dividing them by four (Thielsch and Moshagen, 2015). The internal consistency of the VisAWI total score was excellent according to George and Mallery (2019) [α = 0.96, 95% CI (0.95, 0.97), ω_h = 0.95, 95% CI (0.93, 0.96)], and between good and excellent for the four subscales: *Simplicity* with five items [α = 0.86, 95% CI (0.83, 0.89), ω = 0.86, 95% CI (0.82, 0.88)], *diversity* with five items [α = 0.87, 95% CI (0.84, 0.90), ω = 0.88, 95% CI (0.84, 0.90)], *colorfulness* with four items [α = 0.91, 95% CI (0.89, 0.93), ω = 0.91, 95% CI (0.89, 0.93)], and *craftsmanship* with four items [α = 0.87, 95% CI (0.84, 0.90), ω = 0.87, 95% CI (0.83, 0.90)].

The theoretical structure of the VisAWI was assessed with a Confirmatory Factor Analysis (CFA) using the lavaan package for R (version 0.6-11, Rosseel, 2012). We examined the proposed four-factor model (i.e., simplicity, diversity, colorfulness, and craftsmanship), including a higher-order factor for overall aesthetics. All items were specified to load on their designated factor, and the first item's loading was constrained to one. Multivariate normality was not given (Henze-Zirkler Test = 2.44, $p < 0.0001$); therefore, a robust maximum likelihood estimation method with Huber-White standard errors and a Yuan-Bentler based test statistic was used. Results of the CFA including all 18 items suggested that the proposed model does not adequately fit the data [$\chi^2(131) = 674.47$, $p < 0.0001$, $CFI = 0.84$, $SRMR = 0.08$, $RMSEA = 0.14$].[11] We consequently performed an exploratory factor analysis (EFA) for the VisAWI data, which suggested a two-factor solution. Factor one consisted of the ten positively formulated items of the VisAWI, while the eight negatively formulated items mostly loaded onto the second factor or cross-loaded onto both. It thus appeared that the item wording (positive or negative) influenced the scale's factor structure. Such a phenomenon has been reported for other scales, including the System Usability Scale (SUS) (Brooke, 1996). In the case of the SUS, Lewis and Sauro (2017) recommended treating the scale as a unidimensional measure due to the limited interest that comes with a distinction based on negative/positive item tone. Following this example, we decided to stick with a one-factor solution for the VisAWI as an indicator of *perceived aesthetics* because a distinction between the two factors was theoretically non-sensible. We further refrained from interpreting the four sub-scales of the VisAWI. A one-factor EFA showed that this one-factor solution explained 60% of variance, while a one-factor CFA indicated a comparable fit to the original model [$\chi^2(135) = 728.46$, $p < 0.0001$, $CFI = 0.82$, $SRMR = 0.08$, $RMSEA = 0.15$].

3.3.2. Perceived usability: the UMUX

The UMUX (Finstad, 2010) was used to measure participants' perceived usability of the respective app variant. The UMUX consists of four items rated using a 7-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). The even items of the scale were reversed before scoring, after which responses were transformed into a score ranging from 0 to 100. The survey scale exhibited acceptable internal consistency according to George and Mallery (2019) [α = 0.81, 95% CI (0.76, 0.85), ω = 0.79, 95% CI (0.67, 0.83)].

As with the VisAWI, we performed a CFA to assess the factor structure of the UMUX data as an indicator of scale validity. All four items of the UMUX were specified to load onto one factor, and the loading of the first item was constrained to one. Multivariate normality was again not given (Henze-Zirkler Test = 16.77, $p < 0.0001$); therefore, the same robust maximum likelihood estimation method was used. Results of the CFA suggested an inadequate fit of the proposed model to the data [$\chi^2(2) = 87.52$, $p < 0.0001$, $CFI = 0.73$, $SRMR = 0.14$, $RMSEA = 0.50$]. As with the VisAWI, we thus performed an EFA for the UMUX data. The EFA suggested a two-factor solution, with one factor for the two positively formulated items and a second for the two negative items. Following the same logic as with the VisAWI, we decided to adhere to the originally proposed one-factor solution for the UMUX, representing *perceived usability*, able to explain 52% of variance (according to a one-factor EFA).

3.3.3. Dependent variable: performance score

Following prior research (Moshagen et al., 2009; Sonderegger et al., 2014; Thielsch et al., 2019b), performance was measured both by a *performance score* using six content-related questions and the task completion time for answering these six questions, hereafter referred to as *performance time*. A high performance thus meant answering as many questions of the information foraging task correctly and having a short performance time.

Participants were asked to answer six questions targeting the app's content to assess the performance score (e.g., "Since when has the Master Events agency been in business?"). These questions were developed in iterative discussions with members of the authors' research group. The exact questions are documented in

[10] https://osf.io/amvsk

[11] *CFI* = Comparative Fit Index; *SRMR* = Standardized Root Mean Square Residual; *RMSEA* = Root Mean Square Error of Approximation. The following criteria were seen as an indication of good model fit: Low χ^2 value and $p >$ 0.05 for the Chi-squared test, $RMSEA < 0.06$, $SRMR \leq 0.08$ and $0.95 \leq CFI \leq 1$ (Hu and Bentler, 1999).

the Supplementary tables and figures on OSF. Each question asked for specific details about the fictional event agency and offered four answer choices, of which only one was correct. Participants had to select the correct answer in each case. Answers to the questions were presented in randomized order to avoid any order effects. One point was awarded for each correct answer, resulting in a minimum of 0 and a maximum of 6 points per participant. The score obtained represented the performance score. The average performance score achieved by participants was 5.10 points ($SD = 1.42$, $range = 0 - 6$). Internal consistency for the six questions was acceptable according to George and Mallery (2019) [$\alpha = 0.76$, 95% CI (0.70, 0.82), $\omega = 0.77$, 95% CI (0.71, 0.83)]. In addition, each item's difficulty and item discrimination was considered to evaluate the performance score further. The mean value across all respondents for each item served as item difficulty, indicating how many participants answered the item correctly. Item difficulty ranged from 0.74 to 0.95, indicating that all items had a reasonable and comparable level of difficulty and could thus be mastered by conscientious participants, although the items were arguably on the easier side. This is comparable to past research, where most participants were able to complete the performance tasks [82% successful task completion in Sonderegger et al. (2014) and difficulty of 0.76 in Thielsch et al. (2019a)]. Item discriminatory power was calculated from the correlation of the item with the score across the other five performance questions (corrected item-total correlation). Values ranged from 0.52 to 0.66, all within the ideal range of between 0.40 and 0.70 (Moosbrugger and Kelava, 2000) and above the lowest acceptable discriminatory power of 0.30 according to Borg and Groenen (2005). The Supplementary tables and figures on OSF contain all values for item difficulty and discriminatory power.

Finally, we conducted a CFA to assess the factor structure of the performance items. All six performance questions were specified to load onto one factor, and the loading of the first item was constrained to one. The same robust maximum likelihood estimation method was used as multivariate normality was again not given (Henze-Zirkler Test $= 90.55$, $p < 0.0001$). Results of the CFA mostly suggested that the proposed model adequately fits the data [$\chi^2(9) = 14.91$, $p = 0.09$, $CFI = 0.97$, $SRMR = 0.04$, $RMSEA = 0.07$]. Only the RMSEA was slightly above the desired value of < 0.06 (Hu and Bentler, 1999).

3.3.4. Dependent variable: performance time

Performance time was collected automatically by the online survey tool. The average time needed by participants to answer all six questions was 2.59 minutes ($SD = 2.15$ minutes, $range = 0.22 - 14.07$ minutes).

3.4. Procedure

The online study featured a between-subjects design with manipulated app aesthetics (high vs. low). Participants were randomly assigned to one of two conditions, resulting in two groups of comparable size (high aesthetics: $n = 139$; low aesthetics: $n = 142$). The two groups did not differ significantly regarding the demographic variables age [$F_{(1,279)} = 0.34$, $p = 0.56$, $\eta^2 < 0.01$.] and gender [$\chi^2(3) = 3.30$, $p = 0.35$, Cramer's $V = 0.11$]. The study consisted of four phases and took participants on average 8.94 minutes to complete ($SD = 3.59$ min, $range = 2.05 - 18.98$ minutes). Data collection for the study was conducted using the online survey tool Unipark.[12]

In the study's first phase, the survey platform automatically checked if participants accessed the study using a mobile device. Access from other device types was denied. Once participants could access the site, they were presented with an introduction briefly explaining the study's purpose. Here, participants were informed about the study characteristics (duration of data storage, anonymity, and compensation) and provided informed consent. Afterward, demographic data (age and gender) was collected. Participants had to be at least 18 years old to participate. Finally, participants were asked whether they were affected by visual or color impairments to ensure they could perceive all aspects of the aesthetic manipulation.

In the second phase, participants were presented with a cover story and a detailed task description (exact wording provided in the Supplementary tables and figures on OSF). Next, participants were randomly assigned to the aesthetic or unaesthetic variant of the app. As a cover story, participants were asked to interact with the web app and review it as part of a usability test, likewise to past research (Hamborg et al., 2014). They were also told that they would have to answer a series of questions about the app's content once they completed their exploration. Here, it was emphasized that a conscientious exploration of the app was necessary to answer the upcoming questions correctly and that they were not allowed to leave the app open while answering the questions. Thus, they received clear goals to fulfill during their interaction with the app (i.e., searching for information on the stimuli website to answer the content questions). By clicking a button, participants were redirected to the app in a new web browser tab and could interact with it at their discretion. It was up to them to decide when to end the exploration and return to the study.

In the third phase of the study, participants answered the six performance questions previously described. Performance time was collected automatically during this process. Afterward, participants filled out the VisAWI and UMUX. The items of each survey scale were presented in randomized order. An attention check item was added among the VisAWI items to ensure adequate data quality ("This is a question to test if you are attentive. Please select (7) strongly agree"). Finally, participants were asked to self-report the quality of their data ("In your honest opinion, did you fill out the survey attentively and should we use your data in our analyses in this study").

In the final phase of the study, participants had the opportunity to provide feedback regarding the survey. Afterward, they received a personalized completion code to claim their compensation through MTurk and were debriefed on the study's purpose. Participants received $2 upon full completion of the study. The OSF

[12] https://www.unipark.com

TABLE 1 Mean, standard deviation and range for key variables sorted by app variant (aesthetic vs. unaesthetic).

	Aesthetic (*n* = 139)			Unaesthetic (*n* = 142)		
	Mean	SD	Range	Mean	SD	Range
VisAWI—Simplicity	5.67	1.03	2.80–7.00	4.27	1.45	1.20–7.00
VisAWI—Diversity	5.31	1.05	2.20–7.00	4.06	1.62	1.20–7.00
VisAWI—Colorfulness	5.81	1.06	2.00–7.00	3.67	1.83	1.00–7.00
VisAWI—Craftsmanship	5.68	1.13	1.75–7.00	3.99	1.73	1.00–7.00
VisAWI—Total Score	5.62	0.95	2.94–7.00	4.00	1.54	1.23–7.00
UMUX score	80.19	18.47	25.00–100.00	61.44	24.43	4.17–100.00
Performance time (minutes)	2.52	2.03	0.22–14.07	2.65	2.27	0.28–13.32
Performance score	5.26	1.23	0.00–6.00	4.95	1.58	0.00–6.00

SD = standard deviation.

TABLE 2 Results from statistical tests used to compare the two app variants.

Variable investigated	Test used	Test statistics
Perceived aesthetics	Welch's two-sided *t*-test	$t_{(236.20)} = 10.63, p < 0.0001, d = 1.26$
Perceived aesthetics	Wilcoxon rank sum test	$W = 15,877, p < 0.0001$
Perceived usability	Welch's two-sided *t*-test	$t_{(262.33)} = 7.26, p < 0.0001, d = 0.86$
Perceived usability	Wilcoxon rank sum test	$W = 14,260, p < 0.0001$
Performance time	Two-sided *t*-test	$t_{(279)} = -0.52, p = 0.60, d = -0.06$
Performance time	Wilcoxon rank sum test	$W = 9,744.5, p = 0.86$
Performance time	Equivalence test	$t_{(276.74)} = -0.10, p = 0.54$
Performance score	Two-sided *t*-test	$t_{(279)} = 1.82, p = 0.07, d = 0.22$
Performance score	Wilcoxon rank sum test	$W = 10,526, p = 0.28$
Performance score	Equivalence test	$t_{(265.79)} = 0.99, p = 0.84$

d = Cohen's d for effect size.

repository contains a schematic representation of the study process and a printout of the online survey.[13]

4. Results

All analyses were performed using the statistical software R (version 4.2.0, R Core Team, 2022). The level of statistical significance was set at $\alpha = 0.05$. To investigate possible differences between conditions, we used parametric and non-parametric statistical tests of significance. In case of non-significant results, we further used equivalency tests. In addition, we used bootstrapping to gain further insight into the robustness of our findings. For this, we drew 1,000 data sets from our original data (with replacement), sampling the same amount of participants per condition as in the original data ($n_{aesthetic} = 139$, $n_{unaesthetic} = 142$). We then calculated t-tests for each of the 1,000 data sets. Exact means and standard deviations for all key variables per condition are presented in Table 1, and results from the statistical tests are listed in Table 2.

4.1. Manipulation check: perceived aesthetics

First, the subjective aesthetic perception of the two app versions was investigated using the VisAWI data. This was also seen as a manipulation check, examining whether the participants perceived the aesthetics of the two app variants as intended. Using a Welch's two-sided t-test with unequal variances, the aesthetic variant scored significantly higher in the VisAWI total score than the unaesthetic variant. Given the sufficiently large sample size, the t-test should still provide reliable results despite a non-normal distribution of the data (Lumley et al., 2002; Bortz and Schuster, 2010). Nevertheless, a Wilcoxon rank sum test was also calculated because equal variances and normal distribution were not given, showing a significant difference between the two groups. Furthermore, the VisAWI total score of the aesthetic variant exceeded the cut-off for an aesthetic interface of 4.5 by Hirschfeld and Thielsch (2015), whereas the unaesthetic variant fell below it. Bootstrapping results showed average values of $t = 10.72$ and $p < 0.0001$, with all 1,000 t-tests showing a $p < 0.05$. Out of the 1,000 bootstrapped p-values, 527 were equal to or smaller than the value observed with the actual data. Based on these results, we concluded that the manipulation of

[13] https://osf.io/udjkm

app aesthetics was successful, given that participants perceived the aesthetic app variant as more aesthetic than the unaesthetic one.

4.2. Perceived usability

As discussed in the methods section, only the aesthetics of the two variants of the app were manipulated. Care was taken to keep all other aspects of the apps the same, including avoiding strong manipulations of system usability that have been used in previous studies. Nevertheless, it was expected that users of the aesthetically pleasing variant of the app would exhibit higher levels of subjective usability than users of the unaesthetic one (H1). A comparison of the UMUX ratings for the two variants, using a Welch's two-sided t-test with unequal variances, showed that subjective usability was rated significantly different depending on the app's aesthetics. Users of the aesthetic app rated usability significantly higher than those of the unaesthetic variant. A Wilcoxon rank sum test was also calculated because equal variances and normal distribution were not given, showing a significant difference between the two groups. Bootstrapping results for the UMUX showed average values of $t = 7.36$ and $p < 0.0001$, with all 1,000 t-tests showing a $p < 0.05$, and 522 p-values smaller than or equal to the originally observed value. Results thus favor a robust difference between the two app variants across the 1,000 data sets. This close link between the subjective judgment of aesthetics and perceived usability is consistent with findings from past research (Gu et al., 2016; Minge and Thüring, 2018; Otten et al., 2020) and is in favor of the first hypothesis.

4.3. Task performance

The dependent variable performance was operationalized by task performance time and performance score, which we treated separately in the analysis.

4.3.1. Performance time

Regarding the task completion time of the performance tasks, a shorter performance time was expected for the aesthetic variant of the app than the unaesthetic one (H2). A comparison of the performance time for the two variants, using a two-sided t-test with equal variances, showed no significant difference between users of the aesthetic app compared to the unaesthetic variant. Because the data were not normally distributed, a Wilcoxon rank sum test was also calculated, showing no significant difference between the two groups.

Given the non-significant difference between the two conditions, we further calculated tests of equivalence (Lakens, 2017; Lakens et al., 2018) to see whether there truly was no meaningful effect or if there was insufficient statistical power to detect the presence or absence of a meaningful effect. Based on the effect from the meta-analysis by Thielsch et al. (2019b, $g = 0.06$), we set the smallest effect size of interest at $d = 0.05$. The equivalence test was non-significant, thus the two groups could not be considered statistically equal. Finally, bootstrapping results showed average values of $t = -0.47$ and $p = 0.46$, with 933 out of 1,000 t-tests non-significant and no p-values smaller than or equal to the observed value. From this, we concluded that the groups did not differ significantly regarding the performance time but were also statistically non-equivalent. These results, therefore, argue against the second hypothesis, considering descriptive statistics, the significance tests, and the results from bootstrapping. Only the equivalency test indicated a possible difference.

4.3.2. Performance score

Regarding the performance score, a higher performance score was expected in the aesthetic condition than in the unaesthetic one (H3). A comparison of the performance score for the two variants, using a two-sided t-test with equal variances, showed no significant difference in the performance score between users of the aesthetic app compared to the unaesthetic variant. A Wilcoxon rank sum test was also calculated because normal distribution was not given, which showed no significant difference between the two groups.

Because of the non-significant difference, we again performed an equivalence test with a smallest effect size of interest of $d = 0.10$ based on the effect from Thielsch et al. (2019b, $g = 0.12$). The equivalence test was non-significant, indicating that the performance score for the two groups was not equal. The bootstrapping of 1,000 data sets showed an average of $t = 1.86$ and $p = 0.17$, with 449 significant t-tests and 526 p-values smaller than or equal to the observed value. These results thus provided mixed evidence concerning the third hypothesis that higher app aesthetics improves performance. While results from the t-test and the Wilcoxon rank sum test provided evidence against H3, the equivalence test showed that the two groups were not equivalent concerning the performance score. The bootstrapping further revealed that while the average p-value was not significant, almost half of all bootstrapped t-tests would be (44.90%).

4.4. Correlations among variables

Finally, Pearson's product-moment correlations were calculated to investigate further the relationships among the UMUX score, the VisAWI score, and the performance measures (time and score). Results showed a significant large positive correlation between the UMUX and VisAWI scores [$r(279) = 0.79$, 95% CI (0.74, 0.83), $p < 0.0001$]. There was one additional significant small positive correlations between the performance score and the UMUX score [$r(279) = 0.23$, 95% CI (0.11, 0.33), $p < 0.001$]. All other correlations were non-significant. Table 3 highlights correlations among key variables considered in the present study, and the Supplementary material contain all correlations, including the sub-scales of the VisAWI.

5. Discussion

The idea that aesthetics has a measurable impact on performance has been the focus of numerous research studies (e.g., Douneva et al., 2016; Gu et al., 2016; Thielsch et al., 2019a; Baughan et al., 2020; Reppa et al., 2021), including a meta-analysis (Thielsch et al., 2019b). However, to the extent

TABLE 3 Correlations among key variables investigated.

	VisAWI score	UMUX score	Performance time
UMUX score	0.79****		
Performance time	0.00	−0.10	
Performance score	0.05	0.23***	0.10

****$p < 0.0001$; ***$p < 0.001$.

of our knowledge, little to no empirical evidence for such an effect exists in the context of smartphone devices. Furthermore, there appears to be no other study investigating the impact of aesthetics on performance that worked with a smartphone app whose actual layout was aesthetically manipulated. Therefore, the present study provides empirical evidence for the influence of aesthetics on performance in the context of smartphone use. Following the call from past research (Thielsch et al., 2015, 2019b), great care was taken to develop both a realistic app and a set of performance tasks for participants' interaction. For this purpose, the aesthetics of a smartphone web app were manipulated to develop two aesthetically different variants of an otherwise identical app. In addition, while the performance questions used were relatively easy, favorable CFA results, high internal consistency, and consistent item analysis metrics show that the items formed a uniform performance measure. We validated all study elements in preliminary discussions to ensure a high transferability of results into practice. Results showed that the two app variants significantly differed in participants' perceived usability and perceived visual aesthetics. No statistically significant differences in performance time or performance score were found. However, equivalency tests also showed that the two groups were not statistically equivalent concerning both performance measures. Furthermore, bootstrapped *t*-tests for the performance score were significant around half of the time (44.90%). These results, alongside the slightly higher performance score in the aesthetic condition, thus point towards an effect of app aesthetics on performance.

5.1. Manipulation of app aesthetics

A notable strength of the present study was that the participants interacted with a realistic smartphone web app manipulated in the aesthetics of its user interface. Therefore, participants based their impressions on real interactions rather than mere screenshots or mock-ups. Consequently, the study's effects were found after an actual interaction with a functional smartphone app. The duration of this interaction was not constrained, just as an interaction in everyday life might not be subject to any particular constraints either. To our knowledge, no comparable experimental setup with smartphone apps has been used in past research to study performance in this context. Therefore, the present study extends the existing literature by ensuring that the interaction with a system took place for a longer time and that the system under consideration was an interactive app. This realistic interaction with an app is a crucial addition to the existing literature, as most studies have focused only on screenshots (Thielsch et al., 2015), computer applications (Gu et al., 2016; Otten et al., 2020), or devices manipulated in their external aesthetics rather than the actual interface (Sonderegger and Sauer, 2010; Sonderegger et al., 2014; Minge and Thüring, 2018).

The manipulation of aesthetics used in this study resulted in a significant difference between the two app variants and a large effect of said manipulation on participant's perceived aesthetics ($d = 1.26$, Cohen, 1988). Therefore, the results of this study provide evidence that the chosen manipulation of aesthetics, based on the findings of Seckler et al. (2015) and the definition of aesthetics by Moshagen and Thielsch (2010), is effective in the context of smartphone apps. The present findings further indicate that the results from Seckler et al. (2015) initially found in a desktop computer context are transferable to mobile devices. This effect of the aesthetics manipulation implies that design aesthetics play a similar role in the context of mobile smartphone devices regarding the user's subjective perception of aesthetics compared to desktop computers. Considering that design is constantly evolving, and people's perceptions and tastes change over the years (Ntoulas et al., 2004), the findings from the present study further show that results from several years ago can still be applied to current applications. The present study's findings thereby provide guidance for professionals in research and industry concerning the aesthetics of digital applications.

5.2. Perceived aesthetics and usability

Although we took care to manipulate the two app variants solely in their aesthetics, participants interacting with the aesthetic variant of the app rated it as significantly more usable after the interaction, showing a large effect of the aesthetics manipulation on perceived usability ($d = 0.86$). Thus, results favor the first hypothesis that users of the aesthetic variant experienced significantly higher subjective usability than users of the unaesthetic one (H1). This finding is consistent with past research (Moshagen et al., 2009; Sonderegger and Sauer, 2010; Sonderegger et al., 2014; Gu et al., 2016; Minge and Thüring, 2018; Otten et al., 2020; Schrepp et al., 2021). Consequently, this study provides further evidence for aesthetics' effect on perceived usability, expanding past evidence to the context of smartphone web apps. One explanation for these results is a so-called halo effect of the aesthetics manipulation on perceived usability, which has been discussed in past research (Tractinsky et al., 2000). Applied to the results found here, it postulates that the high aesthetics of the app implies high subjective usability. As a result, the participants perceive higher subjective usability, although both variants are objectively the same. The present study hence provides evidence that such a halo effect between aesthetics and usability exists not only in a desktop computer context but also in the context of smartphones.

5.3. The effect of aesthetics on performance

Numerous studies have already explored the interaction of aesthetics and performance (e.g., Sauer and Sonderegger, 2011; Sonderegger et al., 2014; Douneva et al., 2016; Gu et al., 2016; Thielsch et al., 2019a; Baughan et al., 2020; Reppa et al., 2021). Despite this, there is still no consensus on whether aesthetics affect performance, as research findings so far have been too ambivalent (Thielsch and Niesenhaus, 2017). This is especially the case for smartphone devices, where there is still little to no research that addresses the aesthetics of the actual user interface of smartphone apps and their effects on performance.

5.3.1. Performance time

Concerning the effect of aesthetics on performance time, results did not reveal a significant difference between the two conditions, consequently leading to the rejection of hypothesis two (H2, shorter performance time for the aesthetic app variant compared to the unaesthetic app). While the two groups were also statistically non-equivalent, results from the bootstrapping showed no significant difference in most cases (93.30%). These findings correspond to the results of Thielsch et al. (2019a), who also found no significant effect of aesthetics on performance time. A possible explanation for this non-significant difference could be that participants did not have a time limit to complete their task in the present study. Thus, the factor time might not have been relevant for the participants, leading to an absence of time pressure, causing the app exploration to take about the same amount of time for participants in both conditions. On the other hand, the present study worked with a crowd-sourced sample from MTurk, where participants are likely to be pressured to complete as many tasks in as little time as possible to increase their payment. Therefore, time might have played a similar and essential role for participants in both conditions. Furthermore, there was substantial variability in performance time across participants in both groups. Given that the online survey platform automatically collected the time participants spent on the survey page containing the performance questions, we could not monitor participants' actual behavior during this time. It is thus possible that some participants had the performance questions open during exploration despite instructions telling them not to, leading to a longer performance time. Others who followed the instructions likely had shorter performance times, reflecting the time spent just answering the questions without the exploration. This limitation of the performance time variable has to be kept in mind when interpreting the results, although the issue was presumably present in both conditions.

5.3.2. Task performance

The present work provides mixed evidence concerning the effect of aesthetics on user performance. Using a set of self-developed questions, summarized in a performance score, results revealed a small but non-significant effect of aesthetics on performance ($d = 0.22$), comparable to the effect reported in the meta-analysis by Thielsch et al. (2019b, $g = 0.12$). This agreement regarding a small effect strengthens the assumption that app interface aesthetics affect performance. However, results showed no statistically significant difference between conditions. Still, while we found no significant difference, we also found no statistical equivalence between the two groups. Taken alongside the descriptively higher performance score for the aesthetic condition and the results from bootstrapping, our findings point toward an effect of app aesthetics on user performance. Results thus indicate that participants might perform significantly better with an app's aesthetic variant than with the unaesthetic one, which favors hypothesis three (H3, higher performance expected for users interacting with the aesthetic app compared to the unaesthetic variant).

Several reasons might explain the absence of a statistically significant difference in performance in the present study. First, most participants answered the questions correctly, given the high average performance scores in both conditions. Thus, they might have already had the questions open while exploring the app, despite the instructions telling them otherwise. This behavior might have influenced participants' performance in both conditions, causing performance to be better than initially expected. Second, the combination of both non-significant null hypotheses significance tests and equivalency tests indicates that the study might have been statistically underpowered to investigate the presence or absence of a meaningful effect thoroughly (Lakens et al., 2018). Results from bootstrapping further undermine this point, with around half of all bootstrapped *t*-tests significant. Thus, larger samples are needed in future studies investigating the effect of app aesthetics on performance. Given the limited number of studies on the effects of aesthetics on performance in the smartphone context, the current study's results thus provide initial evidence for this effect. Third, users' motivations also feasibly influence performance. In the present study, completion time likely was more important to participants than correctly following the task instructions and answering the questions, given the crowd-sourced sample. Nevertheless, the fact that most questions were answered correctly by participants in both conditions argues against this assumption. While the present work did not consider users' motivation as a confounding factor for performance, future work should.

The results of the present study suggest that the aesthetics of a web app can affect users' performance to a similar extent as what was previously found in other contexts. Thielsch et al. (2019b) concluded that aesthetics significantly affected performance with mobile devices (e.g., non-smartphone cell phones) and software applications, but not on websites. The present study thus contributes to these findings, showing that app aesthetics has the potential to affect user performance, although further investigation is needed.

5.4. Implications of results

In summary, the present results provide evidence regarding app aesthetics' effect on subjective (perceived aesthetics and usability) and objective (performance time and score) elements of a user's interaction with a smartphone app. While results indicate no or mixed effects on performance, they suggest an apparent effect of aesthetics on users' subjective experience. While such effects have been found in past research, studies in a smartphone context are still limited. The present study thus is among the first to show that close links between objective aesthetics and subjective perceptions of a system exist within a smartphone context. Even if one assumes that aesthetics do not affect performance in a smartphone context, they have apparent effects on the users' subjective perception. Considering that the subjective perception of the app (i.e., aesthetics, usability) differed significantly between conditions, results highlight that while users do not take less time to complete a task with an aesthetic website, they definitely have an improved subjective experience while arguably performing better.

5.4.1. Theoretical explanations

Regarding past explanations from related work, the results do not support any existing ideas concerning aesthetics' effects on performance. For instance, the significantly higher perceived usability and the slightly better performance score in the aesthetic condition speak for the presence of attentional and cognitive effects (Szabo and Kanuka, 1999; Tractinsky et al., 2000). According to this notion, a more aesthetic design would promote the automatic processing of information, thereby increasing performance, which would explain the somewhat better performance score in the aesthetic condition. However, attentional and cognitive effects can not explain why the evaluation of performance time did not reveal any significant differences, given that faster performance times in the aesthetic condition would also be expected. As described above, the halo effect could explain the differences in subjectively perceived higher usability, although performance differences are unrelated to this effect. At the very least, however, it can be stated that the results of this study argue against the prolongation of joyful experience theory (Sonderegger and Sauer, 2010; Sonderegger et al., 2014). The performance times of the two groups did not differ significantly and did not indicate a prolonged exploration of the aesthetic variant, although the MTurk setting likely influenced these results. Therefore, based on the present results, only conjectures can be made regarding theoretical rationales.

5.4.2. How to study performance

The disparate effects of aesthetics on performance highlight the importance of carefully considering how performance can be operationalized. In the present study, we worked with two ways to quantify users' performance: a self-developed set of content-related questions and the time taken to fill out those questions. While the aesthetic manipulation did not affect performance time, we found mixed results for the performance questions, which suggests that aesthetics affect performance differently depending on the chosen performance indicator.

First, this raises the question of what we denote when discussing performance. While completing a task quickly and efficiently might be crucial in some cases, error-free task completion is of greater importance in others. In the present study, our approach focused on the correct gathering of information to answer specific questions while also considering the time taken for this information-gathering. Thus, high performance meant that users processed and recalled information better (i.e., higher performance score) and faster (i.e., shorter performance time). We thus considered performance from two perspectives.

Second, researchers need to think about how they can measure performance. Standardized scales, such as those used for measuring subjective aesthetics and usability, make little sense for performance, given the high context-bound nature of possible tasks. For the present study, we designed questions to measure performance close to real life, but measuring performance has different approaches. In our study, performance was related to the site's content, which is not always the case. Other approaches include the number of errors, number of commands, or the amount of additional information needed for task completion (Thielsch et al., 2019b). When looking at the data from our performance score, we see a ceiling effect, with most participants getting the majority of questions correct. The choice of performance measure thus influenced our results. Different methods for measuring performance will likely highlight different effects that interface aesthetics and other design factors can have on users.

Thus, researchers should consider different ways of operationalizing performance with mobile devices beyond those used in the present study (i.e., number of correct answers, task duration). Future research comparing different performance measures in varying contexts could deliver additional insight into the effects of aesthetics on user performance. Furthermore, the boundaries of these effects should be explored by using a variety of tasks, more questions, or questions with more considerable differences in difficulty.

5.4.3. How to define aesthetics

Another plausible explanation for the disparate results on the relationship between aesthetics and performance, both in the present paper and in past research, is the multi-factorial construct of aesthetics itself. It is conceivable that different facets of aesthetics have distinct effects on performance and therefore require specific explanations for the individual facets. For example, while the color of an app might impact performance, symmetry might not (or vice versa). Within HCI research, there is still no uniform definition of aesthetics, and research studies sometimes show imprecise or even missing definitions of the examined constructs (Thielsch et al., 2019b). A lack of shared definitions complicates the comparability and interpretation of results across research immensely (Flake and Fried, 2020) and could also explain the contradictory results regarding the effect of aesthetics on performance. In the present work, we only had two app variants manipulated in terms of overall aesthetics. App variants with differences in only certain facets of aesthetics could provide further insight. Future research should thus address these questions and investigate the effects of different facets of aesthetics, mentioned in definitions, on performance.

5.4.4. How to measure aesthetics

In line with the question of how to define aesthetics comes the issue of how to measure it. Just as with definitions, there is a lack of common standard regarding how aesthetics is measured (Thielsch et al., 2019b). Hassenzahl and Monk (2010) argued that contradictory results on the effects between usability and aesthetics could be due to different measurement methods. This likely is also the case for aesthetics and performance. Thielsch et al. (2019b) in their meta-analysis looked at methods used to measure aesthetics and found that many researchers rely on unstandardized measures with varying levels of psychometric quality. Furthermore, using unstandardized measures was associated with larger effects than standardized aesthetics measures such as the VisAWI or the scale by Lavie and Tractinsky (2004). Thus, the varying methods used to measure aesthetics further explain the contradictory results in past research. In addition, given that survey scales are based on underlying theoretical models, these models need to be made clear and investigated whenever one uses survey scales for measurement (DeVellis, 2017; Flake and Fried, 2020). However, investigating the factor structure of the VisAWI raised doubt about the current model used for the scale. As briefly mentioned in the methods section, our attempts to confirm the factor structure of the VisAWI were unsuccessful, leading us only to consider the rating of overall perceived aesthetics. These doubts not only limited our possibilities to investigate the effect that different facets of the app aesthetics have on performance but also challenged the underlying theory behind the VisAWI and the understanding of aesthetics by Moshagen and Thielsch (2010). However, neither the theoretical structure nor the psychometric quality of the VisAWI was the focus of the present study. Future research on both the quality of the scales used within aesthetics research and the theoretical models behind them is thus needed.

5.4.5. Practical implications

Past research has shown that aesthetics are a way to stand out in a crowded market, increasing recognition value and thus making pleasing aesthetics a decisive success factor (Bloch et al., 2003; Bhandari et al., 2015, 2019). However, previous work has investigated aesthetics mainly outside the context of smartphone apps. The present study thus extends past findings, showing that users perceive an aesthetic app as more aesthetic and more usable. Furthermore, aesthetics appear to impact user performance, although to a lesser extent. Designers need to be aware of these effects when working on their products. An app with good aesthetics is more attractive to users, possibly causing them to use the app more, even if they perform equally independently of the app's aesthetics. While some have expressed fears in the past regarding a possible negative impact of aesthetics on performance (e.g., Andre and Wickens, 1995), results from the present study further ease these worries. At the very least, aesthetics do not negatively affect user performance but might positively impact it while definitely influencing the user's subjective experience. On top of the effects found in the present study, there are additional consequences of aesthetics already shown in previous studies. Higher user preference, trust, satisfaction, and willingness to reuse are all related to pleasing aesthetics (Moshagen and Thielsch, 2010). Practitioners should always keep this in mind when considering which aspects of software development are most important. Based on this work's findings, it is clear that investing in the design of the interface and placing great emphasis on aesthetic design is worth it.

5.5. Limitations and future research

The first limitation of this work was that the app to interact with was a web app. Using a web app allowed us to distribute our stimuli to participants regardless of their operating system, with no need for participants to install the app. However, differences between web and native apps might have affected the results. Although the editor used to create the app variants was comparable in features and behavior to a native app (Jobe, 2013), readers should note that no native app was used in this study. Future work should thus replicate this study with native apps.

In addition, this study did not collect information about the use context. Several papers (e.g., van Schaik and Ling, 2009; Sonderegger and Sauer, 2010; Iten et al., 2018) mentioned that the positive effect of aesthetics on performance tends to manifest in a work context. Thus, a system's use context may impact the aesthetics' effect on performance, which should be considered in future research. However, because the present study used crowd-sourcing workers, participants were arguably within a work context mindset.

Third, although the present study used an interactive product (instead of just screenshots), the average duration of interaction was still relatively short (given the overall study duration). The present study thus focused mainly on the users' experience during or directly after the interaction while not looking at other relevant time frames, such as the users' experience before the interaction, afterward, or over time. Further investigation during different time points in the users' interaction cycle would allow for a better understanding of whether and how perceived usability and performance change due to interacting with an aesthetically manipulated app and whether the found effects are stable over time.

Fourth, given the MTurk sample, participants were likely not overly interested in exploring the stimuli app in detail but wanted to complete the study as fast as possible. Given that our performance measures were not directly related to workers completing their task on MTurk, and thus receiving their payment, motivation to respond to the performance questions correctly was likely limited. Still, past research has shown that MTurk samples are comparable in quality to other more traditional online samples while demographically more diverse (Buhrmester et al., 2011).

Next, the screening of participants concerning visual and color impairments was based exclusively on self-reporting. It can, therefore, not be ruled out that some participants affected by these types of impairments took part in the study. Future studies should anticipate this and integrate a color and vision test to ensure that all aspects of the aesthetic manipulation are perceived as intended.

Finally, the present paper focused on aesthetics' effects on performance. Therefore, for successful manipulation of aesthetics, the differences between the app's aesthetic and unaesthetic variants were as extensive as possible. Given that the difference in aesthetics between the two variants of the app was rather extreme, future work could look at different levels of aesthetics and find out where the thresholds are for both differences in subjective experience and user performance. Similarly, only two app variants were investigated without detailed differentiation on the level of individual facets of aesthetics. Thus, no conclusions could be drawn as to which facets contributed to the changed performance and perception. Follow-up studies should investigate which aesthetic aspects lead to performance changes, allowing researchers and professionals to draw conclusions for their work and adapt their aesthetic concepts accordingly.

6. Conclusion

The smartphone industry represents a vast market with seemingly endless potential. However, the specifics of smartphone interfaces and their applications have not yet been sufficiently researched to adequately understand user behavior and experience. Specifically, the aesthetics of apps and their effects on users' subjective experience and performance have seen little research in the past. This paper represents a first attempt to investigate the influence of aesthetics on performance in the context of a functional smartphone app. Two variants of a web app were created, manipulated only in terms of aesthetics. Participants in an online study ($N = 281$) were asked to interact with one of the two app variants before answering content-related questions and filling out standardized survey scales on perceived usability and aesthetics. Results showed that the aesthetically pleasing app variant led to a significantly higher perception of aesthetics and usability. Furthermore, the results point toward an effect of aesthetics on performance, with participants interacting with the aesthetic variant exhibiting slightly better performance. Based on this study, it can be concluded that aesthetic smartphone apps not only look nicer but also have the potential to boost performance. Aesthetics is more than just a "nice to have" feature and represents an essential aspect of applications that should always be considered.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author. All supplementary material for this article can be accessed on the Open Science Framework: https://osf.io/qevpk/.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

DU and FB implemented the online study. DU collected the data with the support of SP and wrote the first draft. SP, DU, and FB performed the statistical analysis. SP wrote the second draft of the manuscript. All authors contributed to the conception, design of the study, manuscript revision, read, and approved the submitted version.

Acknowledgments

Special thanks to Nick von Felten and Antony Berbert de Castro Hüsler for their help in preparing this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbas, A., Hirschfeld, G., and Thielsch, M. T. (2022). An arabic version of the visual aesthetics of websites inventory (ar-visawi): Translation and psychometric properties. *Int. J. Hum. Comput. Interact.* 2022, 1–11. doi: 10.1080/10447318.2022.2085409

Accessibility Guidelines Working Group (2018). *Web Content Accessibility Guidelines (WCAG 2.1)*. W3C Web Accessibility Initiative (WAI). Available online at: http://www.w3.org/WAI/intro/wcag (accessed September 9, 2022).

Adepu, S., and Adler, R. F. (2016). A comparison of performance and preference on mobile devices vs. desktop computers. In *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (New York, NY: IEEE), 1–7.

Anderson, M. (2019). *Mobile Technology and Home Broadband 2019*. Pew Research Center. Available online at: https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/ (accessed September 9, 2022).

Andre, A. D., and Wickens, C. D. (1995). When users want what's not best for them. *Ergon. Design* 3, 10–14. doi: 10.1177/106480469500300403

Apple Pty Ltd. (2021). *Further submission in response to the digital platform services inquiry into app marketplaces*. Apple Pty Ltd. Available online at: https://www.accc.gov.au/system/files/Apple%20Pty%20Limited%20%2810%20February%202021%29.pdf (accessed September 9, 2022).

Ater, T. (2017). *Building Progressive Web Apps: Bringing the Power of Native to the Browser*. Newton, MA: O'Reilly Media, Inc.

Bauerly, M., and Liu, Y. (2008). Effects of symmetry and number of compositional elements on interface and design aesthetics. *Int. J. Hum. Comput. Interact.* 24, 275–287. doi: 10.1080/10447310801920508

Baughan, A., August, T., Yamashita, N., and Reinecke, K. (2020). "Keep it simple: How visual complexity and preferences impact search efficiency on websites," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI), 1–10.

Bhandari, U., Chang, K., and Neben, T. (2019). Understanding the impact of perceived visual aesthetics on user evaluations: An emotional perspective. *Inform. Manage.* 56, 85–93. doi: 10.1016/j.im.2018.07.003

Bhandari, U., Neben, T., and Chang, K. (2015). "Understanding visual appeal and quality perceptions of mobile apps: An emotional perspective," in Kurosu, M., editor, *Human-Computer Interaction: Design and Evaluation. HCI 2015. Lecture Notes in Computer Science* (Cham: Springer), 451–459.

Bi, L., Fan, X., and Liu, Y. (2011). Effects of symmetry and number of compositional elements on chinese users' aesthetic ratings of interfaces: Experimental and modeling investigations. *Int. J. Hum. Comput. Interact.* 27, 245–259. doi: 10.1080/10447318.2011.537208

Bloch, P. H., Brunel, F. F., and Arnold, T. J. (2003). Individual differences in the centrality of visual product aesthetics: Concept and measurement. *J. Consum. Res.* 29, 551–565. doi: 10.1086/346250

Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Berlin: Springer Science and Business Media.

Bortz, J., and Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler: Limitierte Sonderausgabe [Statistics for human and social scientists: Limited special edition], 7th Edn*. Berlin, Heidelberg: Springer.

Brooke, J. (1996). Sus: A "quick and dirty" usability scale. *Usabil. Eval. Indus.* 189, 189–194.

Brühlmann, F., Petralito, S., Aeschbach, L. F., and Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods Psychol.* 2, 100022. doi: 10.1016/j.metip.2020.100022

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psycholo. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. New York, NY: Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.

Csikszentmihalhi, M. (1997). *Finding Flow: The Psychology of Engagement With Everyday Life,1st Edn*. New York, NY:Basic Books.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *J. Exp. Soc. Psychol.* 66, 4–19. doi: 10.1016/j.jesp.2015.07.006

Cyr, D., Head, M., and Larios, H. (2010). Colour appeal in website design within and across cultures: A multi-method evaluation. *Int. J. Hum. Comput. Stud.* 68, 1–21. doi: 10.1016/j.ijhcs.2009.08.005

De Angeli, A., Sutcliffe, A., and Hartmann, J. (2006). "Interaction, usability and aesthetics: What influences users' preferences?" in *Proceedings Of The 6th Conference On Designing Interactive Systems* (New York, NY: Association for Computing Machinery), 271–280.

DeVellis, R. F. (2017). *Scale Development: Theory and Applications, 4th Edn*. Thousand Oaks, CA: SAGE publications, Inc..

Dion, K., Berscheid, E., and Walster, E. (1972). What is beautiful is good. *J. Personal. Soc. Psychol.* 24, 285–290. doi: 10.1037/h0033731

Douneva, M., Haines, R., and Thielsch, M. T. (2015). *Effects of Interface Aesthetics on Team Performance in a Virtual Task. ECIS 2015 Research-in-Progress Papers* Münster: Association for Information Systems (AIS), 60.

Douneva, M., Jaron, R., and Thielsch, M. T. (2016). Effects of different website designs on first impressions, aesthetic judgements and memory performance after short presentation. *Interact. Comput.* 28, 552–567. doi: 10.1093/iwc/iwv033

El-Kassas, W. S., Abdullah, B. A., Yousef, A. H., and Wahba, A. M. (2017). Taxonomy of cross-platform mobile applications development approaches. *Ain Shams Eng. J.* 8, 163–190. doi: 10.1016/j.asej.2015.08.004

Finstad, K. (2010). The usability metric for user experience. *Interact. Comput.* 22, 323–327. doi: 10.1016/j.intcom.2010.04.004

Flake, J. K., and Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Adv. Methods Practices Psychol. Sci.* 3, 456–465. doi: 10.1177/2515245920952393

Fortmann-Roe, S. (2013). Effects of hue, saturation, and brightness on color preference in social networks: Gender-based color preference on the social networking site twitter. *Color Res. Appl.* 38, 196–202. doi: 10.1002/col.20734

Furr, M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. London: SAGE Publications, Ltd.

George, D., and Mallery, P. (2019). *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference, 16th Edn*. New York, NY: Routledge.

Groth, A., and Haslwanter, D. (2015). "Perceived usability, attractiveness and intuitiveness of responsive mobile tourism websites: A user experience study," in *Information and Communication Technologies in Tourism 2015*, ed I. Tussyadiah (Cham: Springer International Publishing),593–606.

Gu, H., Hou, W., Qin, X., Zhang, L., and Dai, Y. (2016). "The effects of aesthetics in usability testing for B2C E-commerce websites," in *Proceedings of the Fourth International Symposium on Chinese CHI* (New York, NY: Association for Computing Machinery), 1–5.

Guo, F., Wang, X.-S., Shao, H., Wang, X.-R., and Liu, W.-L. (2020). How users first impression forms on mobile user interface?: An ERPs Study. *Int. J. Hum. Comput. Interact.* 36, 870–880. doi: 10.1080/10447318.2019.1699745

Hamborg, K.-C., Hülsmann, J., and Kaspar, K. (2014). The interplay between usability and aesthetics: More evidence for the "what is usable is beautiful" notion. *Adv. Hum. Comput. Interact.* 2014, 15. doi: 10.1155/2014/946239

Hartmann, J., Sutcliffe, A., and Angeli, A. D. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Trans. Comput. Hum. Interact.* 15, 1–30. doi: 10.1145/1460355.1460357

Hassenzahl, M. (2018). "The Thing and I: Understanding the Relationship Between User and Product," in *Funology 2 - From Usability to Enjoyment*, eds M. Blythe and A. Monk (Cham: Springer), 301–313.

Hassenzahl, M., and Monk, A. (2010). The inference of perceived usability from beauty. *Hum. Comput. Interact.* 25, 235–260. doi: 10.1080/07370024.2010.500139

Hausman, A. V., and Siekpe, J. S. (2009). The effect of web interface features on consumer online purchase intentions. *J. Bus. Res.* 62, 5–13. doi: 10.1016/j.jbusres.2008.01.018

Hirschfeld, G., and Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. *Ergonomics* 58, 310–320. doi: 10.1080/00140139.2014.965228

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Hum. Comput. Stud.* 64, 79–102. doi: 10.1016/j.ijhcs.2005.06.002

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Eq. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., and DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* 27, 99–114.

International Organization for Standardization (2018). *ISO 9241-11:2018 Ergonomics of Human-System Interaction Part 11: Usability: Definitions and Concepts*. Geneva: International Standardization Organization, Vernier.

Iten, G. H., Troendle, A., and Opwis, K. (2018). Aesthetics in context the role of aesthetics and usage mode for a websites success. *Interact. Comput.* 30, 133–149. doi: 10.1093/iwc/iwy002

Jobe, W. (2013). Native apps vs. mobile web apps. *Int. J. Interact. Mob. Technol.* 7, 27–32. doi: 10.3991/ijim.v7i4.3226

Kurosu, M., and Kashimura, K. (1995). "Apparent usability vs. inherent usability: Experimental analysis on the determinants of the apparent usability," in *Conference Companion on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 292–293.

Lai, C.-Y., Chen, P.-H., Shih, S.-W., Liu, Y., and Hong, J.-S. (2010). Computational models and experimental investigations of effects of balance and symmetry on the aesthetics of text-overlaid images. *Int. J. Hum. Comput. Stud.* 68, 41–56. doi: 10.1016/j.ijhcs.2009.08.008

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.* 8, 355–362. doi: 10.1177/1948550617697177

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Adv. Methods Pract. Psychol. Sci.* 1, 259–269. doi: 10.1177/2515245918770963

Lavie, T., and Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *Int. J. Hum. Comput. Stud.* 60, 269–298. doi: 10.1016/j.ijhcs.2003.09.002

Leder, H., Belke, B., Oeberst, A., and Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *Br. J. Psychol.* 95, 489–508. doi: 10.1348/0007126042369811

Lee, S., and Koubek, R. J. (2010). Understanding user preferences based on usability and aesthetics before and after actual use. *Interact. Comput.* 22, 530–543. doi: 10.1016/j.intcom.2010.05.002

Lewis, J. R., and Sauro, J. (2017). Revisiting the factor structure of the system usability scale. *J. Usabil. Stud.* 12, 183–192.

Lindgaard, G. (2007). Aesthetics, visual appeal, usability and user satisfaction: What do the user's eyes tell the user's brain? *Austr. J. Emerg. Technol. Soc.* 5:1–14.

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., and Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Trans. Comput. Hum. Interact.* 18, 1–30. doi: 10.1145/1959022.1959023

Lingelbach, K., Tagalidou, N., Markey, P. S., Föll, B., Peissner, M., and Vukelić, M. (2022). "Examining joy of use and usability during mobile phone interactions within a multimodal methods approach," in *Proceedings of Mensch Und Computer 2022, MuC '22* (New York, NY: Association for Computing Machinery), 276–285.

Liu, Y., Liu, X., Ma, Y., Liu, Y., Zheng, Z., Huang, G., et al. (2015). "Characterizing RESTful web services usage on smartphones: A tale of native apps and web apps," in *2015 IEEE International Conference on Web Services* (New York, NY: IEEE), 337–344.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Ann. Rev. Publ. Health* 23, 151–169. doi: 10.1146/annurev.publhealth.23.100901.140546

Ma, Y., Liu, X., Liu, Y., Liu, Y., and Huang, G. (2017). A tale of two fashions: An empirical study on the performance of native apps and web apps on android. *IEEE Trans. Mob. Comput.* 17, 990–1003. doi: 10.1109/TMC.2017.2756633

McDonald, R. P. (1999). *Test Theory: A Unified Treatment, 1st Edn*. New York, NY: Psychology Press.

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Michailidou, E., Harper, S., and Bechhofer, S. (2008). "Visual complexity and aesthetic perception of web pages," in *Proceedings of the 26th Annual ACM International Conference on Design of Communication* (New York, NY: Association for Computing Machinery), 215–224.

Minge, M., and Thüring, M. (2018). Hedonic and pragmatic halo effects at early stages of user experience. *Int. J. Hum. Comput. Stud.* 109, 13–25. doi: 10.1016/j.ijhcs.2017.07.007

Moosbrugger, H., and Kelava, A. (2000). *Testtheorie und Fragebogenkonstruktion [Test Theory and Questionnaire Construction], 3rd Edn*. Berlin; Heidelberg: Springer.

Moshagen, M., Musch, J., and Göritz, A. S. (2009). A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics* 52, 1311–1320. doi: 10.1080/00140130903061717

Moshagen, M., and Thielsch, M. (2013). A short version of the visual aesthetics of websites inventory. *Behav. Inform. Technol.* 32, 1305–1311. doi: 10.1080/0144929X.2012.694910

Moshagen, M., and Thielsch, M. T. (2010). Facets of visual aesthetics. *Int. J. Hum. Comput. Stud.* 68, 689–709. doi: 10.1016/j.ijhcs.2010.05.006

Nielsen, J., and Budiu, R. (2013). *Mobile Usability*. Berkeley, CA: New Riders.

Ntoulas, A., Cho, J., and Olston, C. (2004). "What's new on the web? the evolution of the web from a search engine perspective," in *Proceedings of the 13th International Conference on World Wide Web* (New York, NY: Association for Computing Machinery), 1–12.

Olivia, A., Mack, M. L., Shrestha, M., and Peeper, A. (2004). "Identifying the perceptual dimensions of visual complexity of scenes," in *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 26* (Seattle, WA: Cognitive Science Society), 1041–1046.

Otten, R., Schrepp, M., and Thomaschewski, J. (2020). "Visual clarity as mediator between usability and aesthetics," in *Proceedings of the Conference on Mensch und Computer* (New York, NY: Association for Computing Machinery), 11–15.

Oyibo, K., and Vassileva, J. (2020). The effect of layout and colour temperature on the perception of tourism websites for mobile devices. *Multimodal Technol. Interact.* 4, 10008. doi: 10.3390/mti4010008

Palmer, S. E., and Schloss, K. B. (2010). "Human preference for individual colors," in *Human Vision And Electronic Imaging XV, Vol. 7527* (Springfield, VA: International Society for Optics and Photonics, Society for Imaging Science and Technology), 752718.

Palmer, S. E., Schloss, K. B., and Sammartino, J. (2013). Visual aesthetics and human preference. *Ann. Rev. Psychol.* 64, 77–107. doi: 10.1146/annurev-psych-120710-100504

Postrel, V. (2004). *The Substance of Style: How the Rise of Aesthetic Value is Remaking Commerce, Culture, and Consciousness, 1st Edn*. New York, NY: HarperCollins Publisher Inc.

Quinn, J. M., and Tran, T. Q. (2010). "Attractive phones don't have to work better: Independent effects of attractiveness, effectiveness, and efficiency on perceived usability," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 353–362.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., et al. (2013). "Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 2049–2058.

Reppa, I., and McDougall, S. (2015). When the going gets tough the beautiful get going: Aesthetic appeal facilitates task performance. *Psychon. Bullet. Rev.* 22, 1243–1254. doi: 10.3758/s13423-014-0794-z

Reppa, I., McDougall, S., Sonderegger, A., and Schmidt, W. C. (2021). Mood moderates the effect of aesthetic appeal on performance. *Cogn. Emot.* 35, 15–29. doi: 10.1080/02699931.2020.1800446

Reppa, I., Playfoot, D., and McDougall, S. J. (2008). Visual aesthetic appeal speeds processing of complex but not simple icons. *Proc. Hum. Fact. Ergon. Soc. Ann. Meet.* 52, 1155–1159. doi: 10.1177/154193120805201801

Riegler, A., and Holzmann, C. (2018). Measuring visual user interface complexity of mobile applications with metrics. *Interact. Comput.* 30, 207–223. doi: 10.1093/iwc/iwy008

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.6–11. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Salimun, C., Purchase, H. C., Simmons, D. R., and Brewster, S. (2010). "The effect of aesthetically pleasing composition on visual search performance," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (New York, NY: Association for Computing Machinery), 422–431.

Sauer, J., and Sonderegger, A. (2011). The influence of product aesthetics and user state in usability testing. *Behav. Inform. Technol.* 30, 787–796. doi: 10.1080/0144929X.2010.503352

Schrepp, M., Otten, R., Blum, K., and Thomaschewski, J. (2021). What causes the dependency between perceived aesthetics and perceived usability? *Int. J. Interact. Multimedia Artif. Intell.* 6, 78–85. doi: 10.9781/ijimai.2020.12.005

Seckler, M., Opwis, K., and Tuch, A. N. (2015). Linking objective design factors with subjective aesthetics: An experimental study on how structure and color of websites affect the facets of users visual aesthetic perception. *Comput. Hum. Behav.* 49, 375–389. doi: 10.1016/j.chb.2015.02.056

Smith, A. R. (1978). "Color gamut transform pairs," in *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY: Association for Computing Machinery), 12–19.

Sonderegger, A., and Sauer, J. (2010). The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Appl. Ergon.* 41, 403–410. doi: 10.1016/j.apergo.2009.09.002

Sonderegger, A., Uebelbacher, A., Pugliese, M., and Sauer, J. (2014). "The influence of aesthetics in usability testing: The case of dual-domain products," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 21–30.

Statista Research Department (2022). *Anteil mobiler Endgeräte an allen Seitenaufrufen nach Regionen weltweit im Jahr 2021 [Share of Mobile Devices in All Page Visits by Region Worldwide in 2021]*. Statista. Available online at: https://de.statista.com/statistik/daten/studie/217457/umfrage/anteil-mobiler-endgeraete-an-allen-seitenaufrufen-weltweit/ (accessed September 9, 2022).

Szabo, M., and Kanuka, H. (1999). Effects of violating screen design principles of balance, unity, and focus on recall learning, study time, and completion rates. *J. Educ. Multimedia Hypermedia* 8, 23–42.

Tenzer, F. (2022). *Anzahl der Smartphone-Nutzer weltweit von 2016 bis 2020 und Prognose bis 2024 [Number of Smartphone Users Worldwide From 2016 to 2020 and Forecast to 2024]*. Statista. Available online at: https://de.statista.com/statistik/daten/studie/309656/umfrage/prognose-zur-anzahl-der-smartphone-nutzer-weltweit/ (accessed September 9, 2022).

Thielsch, M. T., Blotenberg, I., and Jaron, R. (2014). User evaluation of websites: From first impression to recommendation. *Interact. Comput.* 26, 89–102. doi: 10.1093/iwc/iwt033

Thielsch, M. T., Engel, R., and Hirschfeld, G. (2015). Expected usability is not a valid indicator of experienced usability. *PeerJ Comput. Sci.* 1, e19. doi: 10.7717/peerj-cs.19

Thielsch, M. T., Haines, R., and Flacke, L. (2019a). Experimental investigation on the effects of website aesthetics on user performance in different virtual tasks. *PeerJ* 7, e6516. doi: 10.7717/peerj.6516

Thielsch, M. T., and Moshagen, M. (2015). *VisAWI Manual*. Ludwigsburg: User Interface Design GmbH.

Thielsch, M. T., and Niesenhaus, J. (2017). *User Experience, Gamification, and Performance, Chapter 5* (Hoboken, NJ: John Wiley & Sons, Ltd), 79–101.

Thielsch, M. T., Scharfen, J., Masoudi, E., and Reuter, M. (2019b). "Visual aesthetics and performance: A first meta-analysis," in *Proceedings of Mensch Und Computer 2019* (New York, NY: Association for Computing Machinery), 199–210.

Thorndike, E. L. (1920). A constant error in psychological ratings. *J. Appl. Psychol.* 4, 25–29. doi: 10.1037/h0071663

Thüring, M., and Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *Int. J. Psychol.* 42, 253–264. doi: 10.1080/00207590701396674

Tractinsky, N., and Hassenzahl, M. (2005). Arguing for aesthetics in human-computer interaction. *i-com* 4, 66–68. doi: 10.1524/icom.2005.4.3.66

Tractinsky, N., Katz, A. S., and Ikar, D. (2000). What is beautiful is usable. *Interact. Comput.* 13, 127–145. doi: 10.1016/S0953-5438(00)00031-X

Tseng, P. Y., and Lee, S. F. (2019). "The impact of web visual aesthetics on purchase intention," in *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)* (Piscataway Township, NJ: Institute of Electrical and Electronics Engineers), 28–31.

Tuch, A. N., Bargas-Avila, J. A., and Opwis, K. (2010). Symmetry and aesthetics in website design: Its a mans business. *Comput. Hum. Behav.* 26, 1831–1837. doi: 10.1016/j.chb.2010.07.016

Tuch, A. N., Bargas-Avila, J. A., Opwis, K., and Wilhelm, F. H. (2009). Visual complexity of websites: Effects on users experience, physiology, performance, and memory. *Int. J. Hum. Comput. Stud.* 67, 703–715. doi: 10.1016/j.ijhcs.2009.04.002

Tuch, A. N., Presslaber, E. E., Stöcklin, M., Opwis, K., and Bargas-Avila, J. A. (2012a). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *Int. J. Hum. Comput. Stud.* 70, 794–811. doi: 10.1016/j.ijhcs.2012.06.003

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., and Bargas-Avila, J. A. (2012b). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in hci. *Comput. Hum. Behav.* 28, 1596–1607. doi: 10.1016/j.chb.2012.03.024

van Schaik, P., and Ling, J. (2009). The role of context in perceptions of the aesthetics of web pages over time. *Int. J. Hum. Comput. Stud.* 67, 79–89. doi: 10.1016/j.ijhcs.2008.09.012

Wiecek, A., Wentzel, D., and Landwehr, J. R. (2019). The aesthetic fidelity effect. *Int. J. Res. Market.* 36, 542–557. doi: 10.1016/j.ijresmar.2019.03.002

Xing, J., and Manning, C. A. (2005). *Complexity and Automation Displays of Air Traffic Control: Literature Review and Analysis, Technical Report*. Washington, DC: US Department of Transportation, Office of Aerospace Medicine.

Zhu, D. H., Deng, Z. Z., and Chang, Y. P. (2020). Understanding the influence of submission devices on online consumer reviews: A comparison between smartphones and PCs. *J. Retail. Consum. Serv.* 54, 102028. doi: 10.1016/j.jretconser.2019.102028

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2023.2258634

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS   Check for updates

Development and Validation of a Positive-Item Version of the Visual Aesthetics of Websites Inventory: The VisAWI-Pos

Sebastian A. C. Perrig, Nick von Felten, Marimo Honda, Klaus Opwis, and Florian Brühlmann

Center for General Psychology and Methodology, University of Basel, Basel, Switzerland

ABSTRACT
Despite its importance, few validated scales exist to measure aesthetics in HCI. One notable exception, the Visual Aesthetics of Websites Inventory (VisAWI), has never been validated in English. Furthermore, the VisAWI contains negatively formulated items, which adversely impact the psychometric quality of survey scales. Consequently, this paper's aim was the development of a positive-item version of the VisAWI, the VisAWI-pos, as a viable alternative to the original scale. Positive alternatives for the negative items of the VisAWI scale were developed and evaluated in a first study ($N = 41$). Afterward, a pre-registered second study ($N = 966$) was conducted to validate the VisAWI-pos. In addition, the VisAWI's English version was formally validated for the first time. Results showed that the English VisAWI has comparable psychometric qualities to validated versions in other languages. Furthermore, the VisAWI-pos provided matching results to the original VisAWI while exhibiting equal or improved psychometric quality.

KEYWORDS
Aesthetics; VisAWI; measurement; questionnaire; survey scale; user experience; usability; websites; reverse-coded items; negative items

1. Introduction

Research in human-computer interaction (HCI) has shown that perceived visual aesthetics is essential to the user experience (UX). Interface aesthetics positively influence, among many things, user satisfaction (Lindgaard, 2007; Seng & Mahmoud, 2020), trustworthiness (Lindgaard et al., 2011; Skulmowski et al., 2016), preference (Lee & Koubek, 2010), and perceived usability (Minge & Thüring, 2018). Despite the importance of aesthetics in HCI research, most HCI researchers investigating aesthetics resort to self-developed scales or single items, while only a few standardized scales see frequent use (Thielsch et al., 2019). Furthermore, few instruments exist to measure perceived aesthetics in the context of websites (Moshagen & Thielsch, 2010). Motivated by the need for proper tools to measure perceived website aesthetics, Moshagen and Thielsch (2010) developed a survey scale, the Visual Aesthetics of Websites Inventory (VisAWI). The VisAWI consists of 18 items across four sub-scales (simplicity, diversity, colorfulness, and craftsmanship). The simplicity facet concerns aspects relevant to an effortless perception and processing of a website's layout. Diversity considers how inventive and dynamic the website's layout is. Items from the colorfulness facet are used to evaluate a website's choice and composition of colors, while craftsmanship asks how skillful a website was designed and if modern technologies were employed appropriately (Moshagen & Thielsch, 2010; Thielsch & Moshagen, 2015). In addition to these four facets, the scale measures a website's overall visual aesthetics, represented in a higher-order factor. A few years after the original VisAWI, Moshagen and Thielsch (2013) developed the VisAWI-S, a four-item short version of the scale. With one item per VisAWI facet, this version offers a way to measure the overall visual aesthetics of a website when resources are too limited for using the full VisAWI. The VisAWI-S consists of four items, one per VisAWI facet (Moshagen & Thielsch, 2013).

Both the VisAWI and VisAWI-S have seen frequent use in UX research over the years[1] and have been translated into multiple languages, including Arabic (Abbas et al., 2022) and Farsi (Saremi et al., 2023). However, given that the studies for the VisAWI and VisAWI-S were conducted with German-speaking participants using original German versions of the scales, the English version has yet to see a formal psychometric evaluation (Abbas et al., 2022). Such evaluation would be crucial given that a scale's psychometric properties can differ substantially between groups, especially across cultural backgrounds (Furr, 2011).

In addition to the unknown psychometric quality of the English VisAWI and VisAWI-S, the VisAWI makes use of reverse-coded items, which cause issues affecting the quality of survey scales, including misresponses by participants (van Sonderen et al., 2013), misunderstandings between researchers and participants (Kam et al., 2021), reduced cross-cultural applicability (Venta et al., 2022), miscodings by researchers (Sauro & Lewis, 2011), lower scale reliability (Suárez Álvarez et al., 2018), and distorted factor structures

CONTACT Sebastian A. C. Perrig sebastian.perrig@unibas.ch Center for General Psychology and Methodology, University of Basel, Missionsstrasse 62a, Basel CH-4055, Switzerland

(Kam & Sun, 2022). Out of the 18 items for the full VisAWI, eight items are reverse-coded (two per sub-scale). Negatively worded or reverse-coded items are items formulated opposite the other scale items' direction, which are then re-coded before analysis. For example, a researcher might formulate an alternative to the item "I like using the system" by replacing "like" with the antonym "dislike", thus creating the reverse-coded item "I dislike using the system." The main idea behind using such items is to avoid biases in participants' answers, namely acquiescence bias (ie, the tendency to agree with most or all items in a scale) and extreme response bias (ie, the tendency to select the extreme answering options of the scale, rather than answers towards the middle), thus reducing measurement error (Dalal & Carter, 2014; Sauro & Lewis, 2011). Numerous research articles have investigated the effects of reverse-coded items on participant responses and the quality of survey scales, highlighting issues caused by those negatively formulated items (see Subsection 2.3 for an overview). Thus, instead of improving response accuracy, reversed items might impair it (Schriesheim & Hill, 1981; van Sonderen et al., 2013). Furthermore, negatively formulated items come with a risk of misinterpretation or mistakes by participants and miscodings by researchers (Sauro & Lewis, 2011). Overall, results from past research show that reverse-coded items are better avoided when working with self-reported survey scales, causing more harm than the biases they are supposed to mitigate. Sauro and Lewis (2011) thus recommended that HCI researchers avoid negative item formulation when developing new scales. Still, many scales, both old and new, within HCI and across other areas of research contain such items, with the VisAWI being only one example. Because of this, Sauro and Lewis (2011) created a positive-worded version of the System Usability Scale (SUS) and demonstrated that this positive-only version of the scale yields comparable results to the original version consisting of both positive and negative items. Kortum et al. (2021) provided additional evidence that the positive SUS performs comparably to the original scale.

In summary, the problems caused by reverse-coded items and the apparent benefits of a positive-item version of a survey scale motivated the development of an alternative version of the VisAWI without negatively worded items. Consequently, the present work aimed to develop a version of the VisAWI which avoids negatively worded items and their associated issues. In addition, the project also formally investigated the psychometric quality of the original English VisAWI for the first time. Based on this, the following two research objectives are addressed:

Objective 1: Creation of a positive-item alternative version of the VisAWI scale (VisAWI-pos) that performs psychometrically at least as well or better than the original scale.

Objective 2: Independent validation of the VisAWI scale in its English version following current psychometric practices.

Thus, an alternative version of the VisAWI was developed and evaluated throughout two studies. The methodology of the two studies was inspired by past work on the VisAWI. Thus, the first study used two manipulated versions of a fictional website, similar to studies six and seven from the original paper on the VisAWI (Moshagen & Thielsch, 2010) and the first study used to develop the VisAWI-S (Moshagen & Thielsch, 2013). In contrast, the second study worked with a set of existing websites, likewise to other work on the VisAWI (Abbas et al., 2022; Moshagen & Thielsch, 2010, 2013; Saremi et al., 2023), to see how the English VisAWI and the newly developed VisAWI-pos would perform using such a setting. Overall, an online study setting was chosen, again comparable to past work on the VisAWI, which allowed for collecting large samples relevant to the psychometric evaluation of the scales. As an initial step, the eight reverse-coded items of the VisAWI were reformulated to create positive alternatives (two to three per negative item). An English language expert then reviewed these alternatives. In the first study, these positive alternatives were employed in an online survey alongside the original VisAWI items to identify those positively worded items psychometrically closest to their original negative counterparts. In study two, the new version of the VisAWI containing only positive items (VisAWI-pos) was evaluated in a pre-registered online survey according to current best practices for examining scale quality. In this online study, the VisAWI-pos was compared with the original English version of the VisAWI, while the psychometric quality of the English VisAWI and the VisAWI-S was simultaneously examined. The present work thus contributes a psychometric investigation into the English versions of the original VisAWI and VisAWI-S, in addition to an alternative version of the scale, which avoids the problematic reverse-coded items.

The remainder of this paper is organized as follows: first, related literature on aesthetics in HCI, the VisAWI, and reverse-coded items is summarized. Next, the development process of the alternative items to the VisAWI's reverse-coded items is described, followed by reporting on the first online study used to select items for the VisAWI-pos. Afterward, the second study is recounted, which was used to investigate the psychometric quality of the newly formed VisAWI-pos alongside the VisAWI and VisAWI-S. Finally, the results of the two studies are discussed, and the implications of those results for using the VisAWI, VisAWI-S, and VisAWI-pos are presented.

2. Literature review

2.1. Aesthetic research in HCI

Aesthetics is among the most frequently measured constructs in UX research (Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018). In the context of the VisAWI, the terms aesthetics and beauty are treated interchangeably and defined "as an immediate pleasurable subjective experience that is directed toward an object and not mediated by intervening reasoning" (Moshagen & Thielsch, 2010, p. 690).

Initial works on aesthetics in HCI included efforts by Kurosu and Kashimura (1995), who showed that the aesthetic appeal of a product strongly affects how users perceive the functional aspects, and thus the usability, of an interface. A few years later, Tractinsky et al. (2000) found strong correlations between perceived aesthetics and perceived usability. Based on their findings, they formulated the notion of "what is beautiful is usable," suggesting that the perceived aesthetics of a system influence users' perceptions of other aspects, such as usability. Much HCI research since then has focused on how aesthetics and usability relate, including work by Tuch et al. (2012), who showed that the effect of aesthetics on usability is reversed under certain conditions (ie, "what is usable is beautiful"). Hassenzahl and Monk (2010) have suggested that contradictory findings concerning the relationship between aesthetics and usability are related to methodological inconsistencies within the studies, such as variations of how aesthetics and usability are measured. In more recent years, Minge and Thüring (2018) have shown that the influence of visual aesthetics on other quality perceptions changes across different stages of the interaction, with visual aesthetics influencing perceived usability at first, while the opposite is true after some interaction (ie, usability impacting visual attractiveness). Furthermore, Schrepp et al. (2021) suggested that visual clarity mediates the relationship between perceived usability and aesthetics, impacting both system perceptions.

Additional findings on aesthetics in HCI have shown that aesthetics influence the overall appeal of systems (Hausman & Siekpe, 2009), users' satisfaction (Lindgaard, 2007; Seng & Mahmoud, 2020), user preference (Lee & Koubek, 2010), intention to use (Pengnate et al., 2019), system trustworthiness (Lindgaard et al., 2011; Skulmowski et al., 2016), and emotion (Bhandari et al., 2019; Seo et al., 2015). Furthermore, numerous research articles have looked at how aesthetics affect not only users' subjective experiences but their performance (for an overview, see Thielsch et al., 2019). Concerning attributes of a system relevant to aesthetics, past work has shown that the complexity and symmetry of a system strongly influence how it is perceived (eg, Seckler et al., 2015). In addition, a website's color can be crucial, with perceived website complexity and colorfulness affecting users' first impressions of website aesthetics (Reinecke et al., 2013) and changes in color resulting in significantly different ratings of website attractiveness (Seckler et al., 2015).

In summary, aesthetics is a crucial factor in users' experiences, interacting with and affecting numerous other aspects of system perception and interaction. Overall, research on aesthetics in HCI has a long tradition and is still ongoing.

2.2. Measuring aesthetics in HCI

Despite the importance of aesthetics in HCI, few instruments exist to measure it (Moshagen & Thielsch, 2010), and most instruments used to measure aesthetics are either self-developed or single-item scales (Abbas et al., 2022; Thielsch et al., 2019). Given that the quality of such scales is often unknown or questionable, psychometrically evaluated tools to measure aesthetics are needed. One of the first attempts to measure aesthetics in the context of HCI was reported by Lavie and Tractinsky (2004), who differentiate between classic and expressive aesthetics. Across four studies, ten items were developed to represent attributes related to how orderly and clear the design of a system is (ie, classic aesthetics), as well as how original and creative it is (ie, expressive aesthetics). Building upon the work by Lavie and Tractinsky (2004), and motivated by a need for precise operational definitions of aesthetics and a well-designed measurement instrument, Moshagen and Thielsch (2010) created a novel visual aesthetics scale, the VisAWI (see Table 1 for an overview of past work on the VisAWI).

The VisAWI was developed and validated across seven studies with a total sample of 2027 German-speaking respondents. Results from these studies showed that the VisAWI was of adequate psychometric quality and that the theoretical model underlying the scale was appropriate. The theoretical model of the VisAWI assumes that the visual aesthetics of a website are represented in one general higher-order factor, which in turn consists of four underlying facets (ie, simplicity, diversity, colorfulness, and craftsmanship). For each of these facets, the scale contains between four and five items, of which two are always negatively formulated (Moshagen & Thielsch, 2010). Moshagen and Thielsch (2013) developed a short version of the VisAWI, the VisAWI-S, across three studies with 1673 German-speaking participants. With one item per facet of the VisAWI, this version of the scale serves as an alternative measure of overall visual aesthetics close to the overall score of the full VisAWI. The authors reported favorable evidence concerning the reliability and validity of the VisAWI-S, but the results were limited to the German version of the scale (Moshagen & Thielsch, 2013). For a more straightforward interpretation of VisAWI results in practice, Hirschfeld and Thielsch (2015) conducted two studies with a total of 972 participants, establishing an optimal cut point for overall VisAWI ratings to be used as a threshold for good aesthetic website design. Recently, two additional versions of the

Table 1. Summary of past work on the VisAWI.

Source	Scale developed	Summary
Moshagen and Thielsch (2010)	VisAWI	Report on developing the original scale across seven studies with 2027 participants, using the German version of the VisAWI. The English translations of the scale items were also documented in this paper.
Moshagen and Thielsch (2013)	VisAWI-S	Developed and validated a short version of the VisAWI-S across three studies with 1673 German-speaking participants.
Hirschfeld and Thielsch (2015)		Conducted two studies with 972 participants to establish a cutoff of 4.5 for overall VisAWI ratings to be used as a threshold of good aesthetics.
Abbas et al. (2022)	AR-VisAWI	Translated the VisAWI into Arabic and showed that this translated version has good psychometric quality in a study with 223 participants.
Saremi et al. (2023)	FV-VisAWI	Created a Farsi version of the VisAWI and investigated its quality in a study with 200 participants, yielding favorable results for this version.

VisAWI in other languages were developed and validated. Abbas et al. (2022) created an Arabic version called the AR-VisAWI with a sample of 223 participants, while Saremi et al. (2023) translated the VisAWI into Farsi and validated the resulting FV-VisAWI with a sample of 200 participants. Both studies showed that the VisAWI could be translated into other languages while retaining its psychometric properties, although certain modifications (eg, item removal) were necessary to achieve those results.

Overall, the VisAWI is a promising scale for measuring aesthetics in the context of HCI, and past evidence has shown that it is of high psychometric quality. However, given that all research to date has focused on versions of the VisAWI in languages other than English, proper evaluation of the English versions of the VisAWI and VisAWI-S is still required.

2.3. The effects of using negatively worded items

A survey scale's items are typically designed to reflect different aspects of a latent construct. Responses to the items are assumed to be affected by this unobservable construct's level, and thus a scale allows researchers to measure the construct indirectly (DeVellis, 2017). A particular type of scale items are reverse-coded or negative items. These items are worded opposite to the direction of the other–positively formulated–scale items and thus need to be re-coded prior to analyzing the responses. The general idea behind using such items is to counteract biases that influence participants' responses to a scale, such as acquiescence bias and extreme response bias, both sources of measurement error (Dalal & Carter, 2014; Sauro & Lewis, 2011). Despite past recommendations to use reverse-coded items in survey scales (eg, Nunnally, 1978), numerous research findings imply that these items cause issues worse than the biases they are supposed to mitigate (see Table 2 for an overview).

First, multiple researchers have shown reverse-coded items to have undesirable effects on a scale's psychometric quality, including its internal consistency and factor structure. Based on a comparison of responses to three versions of the same scale–consisting of only positively formulated items, negative items, or a mix of both–Suárez Álvarez et al. (2018) argued against simultaneously using both item types for four reasons: differences in item comprehension, changes in response variability, worse psychometric properties, and disparities in collected scale scores. Salazar (2015) showed that while using only positive items in a survey might lead to higher acquiescence, adding negative items worsens the situation by substantially reducing the scale's internal consistency and theoretical model fit. Furthermore, distorted factor structures have been reported for multiple scales in psychology (eg, Lindwall et al., 2012; Zhang et al., 2016), demonstrating that reverse-coded items can contaminate a scale's factor structure, making the use of more complex models necessary to achieve good model fit. Within HCI research, Lewis and Sauro (2017) noted an artificial two-factor structure caused by positive/negative item wording for the SUS (Brooke, 1996). Comparable results were also found by Lewis et al. (2013) for the Usability Metric for User Experience (Finstad, 2010) and for a popular scale used to measure trust between people and automated systems (Jian et al., 2000) in the context of artificial intelligence (Perrig, Scharowski, et al., 2023; Scharowski & Perrig, 2023). Thus, reverse-coded items have been shown to distort the factor structure of multiple scales within HCI research, possibly directing researchers toward inaccurate conclusions and complicating the interpretation of data.

In addition to reduced psychometric quality, reverse-coded items can lead to general misunderstandings between participants and researchers or mistakes by participants when filling out a scale (Sauro & Lewis, 2011). Weijters and Baumgartner (2012) argued that negatively worded items sometimes leave too much room for interpretation. Consequently, even cautious participants misunderstand the antonyms used to create the reversed items, leading them to respond contrary to the researchers' intentions. Similarly, Kam et al. (2021) showed that when respondents have a neutral opinion or expression of a construct, they will agree/disagree to a comparable extent with positively and negatively formulated items. While these would be considered misresponses to the survey scale, following the logic behind reverse-coded items, they reflect reasonable answers while showing that these items do not work well for all types of respondents. Additional results by van Sonderen et al. (2013) indicated that reversed items do not prevent inattentive or acquiescent response behavior, rather causing inattention and confusion among participants, increasing respondent mistakes. Participants with lower cognitive abilities also provide more biased responses to reversed items (Gnambs & Schroeders, 2020). Negatively worded items further reduce the cross-cultural applicability of a survey scale, with research results suggesting that positive and negative item wording is interpreted differently across languages and

Table 2. Non-exhaustive summary of past findings on issues related to negatively worded items.

Issue	Source(s)
Lower scale reliability/internal consistency	Barnette (2000); Pilotte and Gable (1990); Salazar (2015); Schriesheim and Hill (1981); Stewart and Frye (2004); Suárez Álvarez et al. (2018)
Distorted factor structures	DiStefano and Motl (2006); Kam and Sun (2022); Lewis and Sauro (2017); Lewis et al. (2013); Lindwall et al. (2012); Perrig, Scharowski, and Brühlmann (2023); Pilotte and Gable (1990); Schmitt and Stuits (1985); Suárez Álvarez et al. (2018); Woods (2006); Zhang et al. (2016)
Misresponses by participants	Gnambs and Schroeders (2020); Sauro and Lewis (2011); van Sonderen et al. (2013); Weijters and Baumgartner (2012)
Misunderstandings between researchers and participants	Kam et al. (2021)
Reduced cross-cultural applicability	Lindwall et al. (2012); Venta et al. (2022); Wong et al. (2003)
Miscodings by researchers	Sauro and Lewis (2011)

cultures (Lindwall et al., 2012; Venta et al., 2022; Wong et al., 2003). Finally, using reverse-coded items can result in miscodings by researchers, such as forgetting to reverse the items before scoring (Sauro & Lewis, 2011).

In summary, there is little evidence that using a mix of positive and negative items solves the problems these items are designed to address, given that their inclusion has a negligible impact on acquiescence bias (van Sonderen et al., 2013). However, more than enough results show that the reverse-coded items themselves negatively impact the survey scales. Furthermore, Dodeen (2023) showed that replacing reverse-coded items with positive alternatives systematically and significantly improved the reliability and factor structure of multiple psychological scales. Motivated by these issues associated with reverse-coded items, Sauro and Lewis (2011) developed a positive version of the SUS, delivering comparable results to the original version of the scale. Given the need for solid instruments to measure aesthetics in HCI, the present work followed their example and set out to create an alternative version of the VisAWI, avoiding negatively formulated items.

3. Item development

As a first step, alternative items to the reverse-scored items of the VisAWI were developed. The goal was to create a collection of appropriate items that retained the meaning of the eight original reverse-scored VisAWI items while avoiding negative wording. To create alternative non-reversed items, the first and third authors formulated positive alternatives to the original negatively worded VisAWI items, first by working independently of each other with multiple dictionaries, then discussing and combining possible alternatives. Survey items are most often reversed in one of two ways, either by negating the target expression (eg, by adding "not") or by using an antonym (eg, "bad" instead of "good") (Suárez Álvarez et al., 2018). Item alternatives thus were either created by removing negations from the original VisAWI items (eg, "The colors do not match" to "The colors match") or by searching for suitable antonyms (eg, "The site appears patchy" to "The site appears uniform"). Next, items were sent to an English language expert offering academic editing for an expert review. The first author then met with the language expert to discuss the items and settle on a selection of 18 items, with two to three positively worded alternatives per original negative VisAWI item. These items can be seen in Table 3 alongside item characteristics from the first study described in the following section.

4. Study one: Item selection

After developing a set of alternatives for the reverse-coded items of the VisAWI, the researchers wanted to refine this item set further, reducing the number of items to ideally one alternative per original reverse-coded VisAWI item. These items could then be combined into a new positive-item version of the VisAWI, which could then be compared to the original. The first study thus aimed to answer the following research question: Which of the newly developed positive items are most suitable as alternatives to the reverse-coded items of the original VisAWI? For this, an online study was conducted where the newly developed alternative items were employed alongside the original VisAWI. The researchers opted for an experimental study with two aesthetically manipulated websites to better compare the original VisAWI ratings with those from the newly developed alternative.

4.1. Methods

A between-subjects online experiment was conducted. Participants were asked to interact with one of two variants of a fictitious event agency website manipulated in terms of aesthetics. After the interaction, they were presented with all items of the VisAWI, both the original and the positive alternatives.

4.1.1. Stimuli

Two versions of the same website were used as stimuli and manipulated regarding the website's aesthetics (aesthetic vs. unaesthetic). The two sites were created as part of another research project (Perrig, Ueffing, et al., 2023) and were mainly manipulated regarding their colorfulness, symmetry, and visual complexity. The choice of colors was derived from past findings regarding color preferences (Seckler et al., 2015), and the manipulation was based on Moshagen and Thielsch (2010)'s conceptualization of aesthetics. Based on informal discussions with four user interface and UX designers, an initial version of the stimuli was designed, which was then manipulated to create seven different variants. To manipulate colorfulness, different color combinations were used, chosen based on past research (Seckler et al., 2015). The different websites' simplicity, or rather the complexity, was manipulated by varying the number of different colors and fonts used throughout the different versions, while symmetry was manipulated by aligning, or not aligning, elements of the website with the site's central vertical axis. After the creation of these first seven stimuli variants, a preliminary evaluation was conducted. Twelve Ph.D. and MSc students enrolled in the HCI program at the authors' university rated screenshots of the seven variants using the VisAWI-S (Moshagen & Thielsch, 2013). In addition, the seven variants were also ranked from most to least aesthetic. Details on this process are reported in Perrig, Ueffing, et al. (2023). The websites rated most and least aesthetic, both with the VisAWI-S and in the rankings, were then used as stimuli in Perrig, Ueffing, et al. (2023) and the present study. To achieve low visual complexity, the aesthetic variant contained only two colors (blue and gray) and only one font type. Symmetry was kept at a maximum, and elements of the site were designed to take up comparable amounts of screen space. In contrast, the unaesthetic website version used six colors and three different fonts to generate high visual complexity. Symmetry was purposefully

Table 3. Descriptive statistics for the original VisAWI items and the proposed alternatives in study one, including correlation between negative items and proposed alternatives (overall and per condition, aesthetic vs. unaesthetic).

		Overall ($N = 41$)		Aesthetic ($n = 22$)		Unaesthetic ($n = 19$)		Correlation with Original Item		
Item Code	Item Wording	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Overall	Aesthetic	Unaesthetic
VisAWI_s1	The layout appears too dense	5.27 (1.52)	1.00–7.00	5.73 (1.42)	3.00–7.00	4.74 (1.49)	1.00–7.00			
VisAWI_s1_pos1	The layout makes good use of white space	4.41 (2.19)	1.00–7.00	5.96 (0.95)	4.00–7.00	2.63 (1.83)	1.00–6.00	0.51***	0.63**	0.33
VisAWI_s1_pos2	**The layout appears clean**	4.49 (2.21)	1.00–7.00	6.14 (0.71)	5.00–7.00	2.58 (1.77)	1.00–6.00	0.53***	0.65**	0.46*
VisAWI_s2	The layout is easy to grasp	5.44 (1.53)	1.00–7.00	6.09 (0.92)	3.00–7.00	4.68 (1.77)	1.00–7.00			
VisAWI_s3	The layout appears well structured	4.83 (2.07)	1.00–7.00	5.91 (1.41)	1.00–7.00	3.58 (2.04)	1.00–6.00			
VisAWI_s4	The site appears patchy	4.44 (2.21)	1.00–7.00	5.64 (1.65)	2.00–7.00	3.05 (1.99)	1.00–7.00			
VisAWI_s4_pos1	**The site appears uniform**	4.88 (1.89)	1.00–7.00	5.86 (1.21)	3.00–7.00	3.74 (1.91)	1.00–7.00	0.51***	0.64**	0.03
VisAWI_s4_pos2	The site appears consistent	5.07 (1.88)	1.00–7.00	6.18 (0.85)	4.00–7.00	3.79 (1.93)	1.00–7.00	0.57***	0.69***	0.16
VisAWI_s5	Everything goes together on this site	4.66 (2.14)	1.00–7.00	6.09 (0.87)	4.00–7.00	3.00 (1.97)	1.00–7.00			
VisAWI_d1	The design is uninteresting	4.54 (2.15)	1.00–7.00	5.23 (1.93)	1.00–7.00	3.74 (2.16)	1.00–7.00			
VisAWI_d1_pos1	**The design is interesting**	4.59 (2.04)	1.00–7.00	5.68 (1.32)	2.00–7.00	3.32 (2.00)	1.00–6.00	0.65***	0.50*	0.66**
VisAWI_d1_pos2	The design is exciting	4.12 (2.18)	1.00–7.00	5.23 (1.77)	1.00–7.00	2.84 (1.92)	1.00–7.00	0.51***	0.24	0.58**
VisAWI_d2	The layout is inventive	4.22 (1.81)	1.00–7.00	5.14 (1.32)	2.00–7.00	3.16 (1.74)	1.00–6.00			
VisAWI_d3	The design appears uninspired	4.17 (2.19)	1.00–7.00	5.41 (1.87)	1.00–7.00	2.74 (1.59)	1.00–7.00			
VisAWI_d3_pos1	**The design appears inspired**	4.44 (1.98)	1.00–7.00	5.68 (1.04)	3.00–7.00	3.00 (1.83)	1.00–7.00	0.75***	0.49*	0.71***
VisAWI_d3_pos2	The design appears innovative	4.22 (2.03)	1.00–7.00	5.46 (1.34)	2.00–7.00	2.79 (1.75)	1.00–6.00	0.48**	0.28	−0.04
VisAWI_d4	The layout appears dynamic	4.37 (1.85)	1.00–7.00	5.27 (1.24)	3.00–7.00	3.32 (1.92)	1.00–6.00			
VisAWI_d5	The layout is pleasantly varied	4.27 (2.16)	1.00–7.00	5.41 (1.56)	1.00–7.00	2.95 (2.01)	1.00–7.00			
VisAWI_col1	The color composition is attractive	4.12 (2.37)	1.00–7.00	5.86 (1.13)	3.00–7.00	2.11 (1.73)	1.00–6.00			
VisAWI_col2	The choice of colors is botched	4.44 (2.35)	1.00–7.00	5.96 (1.33)	3.00–7.00	2.68 (2.03)	1.00–7.00			
VisAWI_col2_pos1	**The choice of colors is perfect**	3.85 (2.36)	1.00–7.00	5.46 (1.50)	2.00–7.00	2.00 (1.73)	1.00–6.00	0.69***	0.27	0.43
VisAWI_col2_pos2	The choice of colors is apt	4.00 (2.19)	1.00–7.00	5.23 (1.77)	1.00–7.00	2.58 (1.74)	1.00–6.00	0.58***	0.11	0.40
VisAWI_col2_pos3	The choice of colors is suitable	4.20 (2.39)	1.00–7.00	5.96 (1.00)	4.00–7.00	2.16 (1.83)	1.00–6.00	0.72***	0.21	0.42
VisAWI_col3	The colors do not match	4.42 (2.40)	1.00–7.00	6.23 (0.92)	4.00–7.00	2.32 (1.77)	1.00–7.00			
VisAWI_col3_pos1	**The colors match**	4.44 (2.29)	1.00–7.00	6.14 (0.83)	4.00–7.00	2.47 (1.81)	1.00–6.00	0.93***	0.76***	0.79***
VisAWI_col3_pos2	The colors fit together	4.24 (2.35)	1.00–7.00	6.14 (0.77)	4.00–7.00	2.05 (1.47)	1.00–6.00	0.95***	0.55**	0.91***
VisAWI_col4	The colors are appealing	4.22 (2.22)	1.00–7.00	5.68 (1.21)	3.00–7.00	2.53 (1.90)	1.00–6.00			
VisAWI_craf1	The layout appears professionally designed	4.07 (2.33)	1.00–7.00	5.73 (1.32)	3.00–7.00	2.16 (1.68)	1.00–6.00			
VisAWI_craf2	The layout is not up-to-date	4.59 (2.13)	1.00–7.00	5.32 (1.84)	1.00–7.00	3.74 (2.18)	1.00–7.00			
VisAWI_craf2_pos1	**The layout is up-to-date**	4.68 (2.10)	1.00–7.00	6.00 (1.27)	2.00–7.00	3.16 (1.83)	1.00–6.00	0.69***	0.53*	0.73***
VisAWI_craf2_pos2	The layout is modern	4.68 (2.09)	1.00–7.00	6.00 (1.07)	3.00–7.00	3.05 (1.87)	1.00–6.00	0.56***	0.51*	0.44
VisAWI_craf3	The site is designed with care	4.68 (2.26)	1.00–7.00	6.23 (0.92)	4.00–7.00	2.90 (2.03)	1.00–6.00			
VisAWI_craf4	The design of the site lacks a concept	4.98 (1.75)	1.00–7.00	5.82 (1.05)	3.00–7.00	4.00 (1.92)	1.00–7.00			
VisAWI_craf4_pos1	**The design of the site has a clear concept**	5.05 (1.87)	1.00–7.00	6.00 (1.07)	4.00–7.00	3.95 (2.01)	1.00–7.00	0.79***	0.47*	0.78***
VisAWI_craf4_pos2	The design of the site has a concept	4.85 (1.82)	1.00–7.00	5.82 (0.91)	4.00–7.00	3.74 (2.00)	1.00–7.00	0.67***	0.41	0.57*
VisAWI_craf4_pos3	The design of the site has an apparent concept.	4.83 (1.76)	1.00–7.00	5.82 (0.80)	4.00–7.00	3.68 (1.89)	1.00–6.00	0.79***	0.70***	0.71***

Correlations calculated with Pearson's r, $* * * = p < 0.001$, $* * = p < 0.01, * = p < 0.05$. Items marked before the study as favorites are presented in bold.

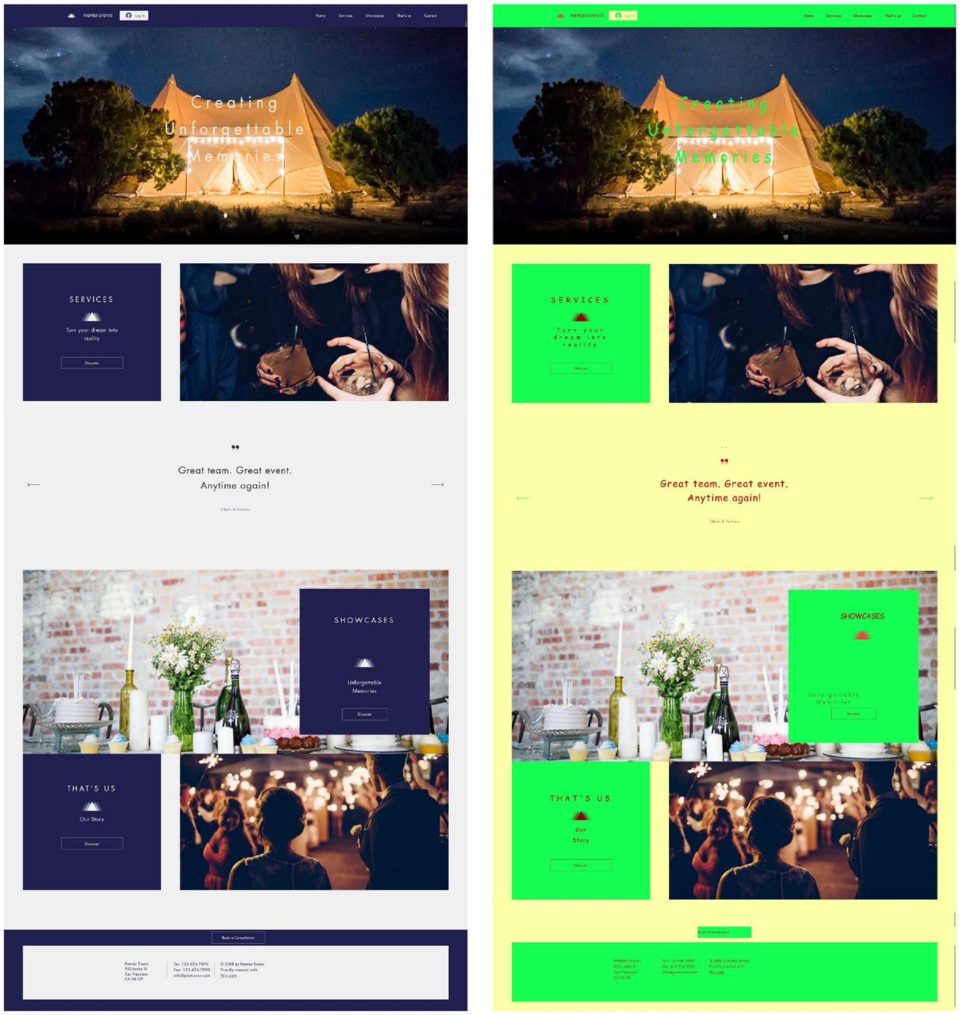

Creating Unforgettable Memories
SERVICES
Turn your dream into reality
Discover
Great team. Great event. Anytime again!
SHOWCASES
Unforgettable Memories
THAT'S US
Our Story
Creating Unforgettable Memories
SERVICES
Turn your dream into reality
Great team. Great event. Anytime again!
SHOWCASES
Unforgettable Memories
THAT'S US
Our Story

Figure 1. Aesthetic (left) and unaesthetic (right) versions of the stimuli website from study one, taken from Perrig, Ueffing, et al. (2023). Note that the first image depicted in the screenshots was replaced with a comparable image for this publication due to copyright.

disregarded wherever possible. Figure 1 contains screenshots of the two websites' landing pages for the desktop version.

4.1.2. Participants

Participants were recruited on Prolific, a crowd-sourcing platform recently shown to have higher data quality than other platforms (Peer et al., 2022). Fifty participants from the United States of America were recruited on Prolific and reimbursed £1.50 for completing the study. Participants were restricted to having no color blindness or visual impairment based on the information given via Prolific. During data cleaning, two observations were removed because they indicated visual impairments, possibly impacting their perception of the manipulated stimuli websites, three because they failed an instructed response item (Curran, 2016), and another three due to a seriousness check (Meade & Craig, 2012). This left a final sample of 41 responses (19 women, 20 men, one self-described, one not specified; mean age 34.20 years, $SD = 12.38$, $\min = 19, \max = 66$).

4.1.3. Measures

The original VisAWI and the alternative positive items were used for data collection. As a task, participants were further asked to respond to a set of questions related to the website content. The content questions and the items of the VisAWI were presented in randomized order.

4.1.3.1. VisAWI. Thirty-six VisAWI items were presented in a randomized order. These included the 18 items from the original VisAWI and the 18 positive alternatives from the item development (two to three per negative original VisAWI item). Responses were collected using a seven-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree).

4.1.3.2. Content questions. Six questions related to the website content were used to ensure that participants interacted with the site. These questions were developed as part of the same research project that the sites themselves were developed for (Perrig, Ueffing, et al., 2023) and are provided in the supplementary materials on OSF (https://osf.io/p84kz).

4.1.4. Procedure

After granting informed consent, participants were randomly assigned to one of two conditions, either interacting with

Table 4. Factor loadings > 0.20 and communalities from study two of all 18 VisAWI-pos items with the four-factor model.

Item name	Item	PA2	PA1	PA3	PA4	h2
VisAWI_s1_pos	The layout appears clean.	0.67				0.67
VisAWI_s2	The layout is easy to grasp.	0.93				0.72
VisAWI_s3	The layout appears well structured.	0.71				0.76
VisAWI_s4_pos	The site appears uniform.	0.62				0.48
VisAWI_s5	Everything goes together on this site.	0.55		0.21		0.71
VisAWI_d1_pos	The design is interesting.		0.74			0.74
VisAWI_d2	The layout is inventive.		0.82			0.63
VisAWI_d3_pos	The design appears inspired.		0.78			0.73
VisAWI_d4	The layout appears dynamic.		0.60			0.59
VisAWI_d5	The layout is pleasantly varied	0.25	0.53			0.65
VisAWI_col1	The color composition is attractive.			0.89		0.83
VisAWI_col2_pos	The choice of colors is perfect.			0.79		0.71
VisAWI_col3_pos	The colors match.	0.23	−0.20	0.63		0.54
VisAWI_col4	The colors are appealing.			0.88		0.81
VisAWI_craf1	The layout appears professionally designed.		0.29		0.45	0.73
VisAWI_craf2_pos	The layout is up-to-date.		0.37		0.38	0.68
VisAWI_craf3	The site is designed with care.	0.35	0.28		0.32	0.75
VisAWI_craf4_pos	The design of the site has a clear concept.	0.57				0.63

Problematic items are marked in bold. PA1–PA4 = factor loadings; h2 = communality.

the aesthetic or unaesthetic version of the website. Participants were further asked to respond to the content questions while keeping the site open. Next, participants responded to the VisAWI items—original and proposed alternatives—followed by demographic questions (age, gender, color blindness, visual impairment). Finally, they had the opportunity to provide feedback before being directed to Prolific for payment. On average, completing the survey took participants 9.06 minutes ($SD = 4.07$, $min = 3.83$, $max = 22.12$).

4.2. Results

Reporting of the results focuses on different item analyses for the VisAWI items and a first comparison of the VisAWI with the newly developed VisAWI-pos. Detailed results can be found on OSF (https://osf.io/cxkme). All results were obtained using the statistical software R (version 4.2.2).

4.2.1. Item analysis and correlations

For the item analysis, descriptive statistics (see Table 3), item difficulty, item variance, discriminatory power, and inter-item correlation computed for all 36 VisAWI items (original and positive alternatives) were considered. No items were flagged as suspicious or removed based on these analyses. Interested readers are referred to OSF for details.

Because the item analysis did not flag any items as overtly suspect and thus provided limited evidence for selecting items for the VisAWI-pos, Pearson's correlations between the original and alternative items were considered to form decisions about the most suitable positive alternatives. Correlations between the reverse-coded VisAWI items and their positive alternatives can be found in Table 3.

4.2.2. Item selection for the VisAWI-pos

Based on the results from the first online study, an initial version of the VisAWI-pos was created. Given that most items performed comparably well in the item analysis, the choice of items primarily relied on correlations and those marked as favorites based on the language center evaluation (see Table 3). For example, correlation results did not strongly favor one alternative simplicity item over another. The item VisAWI_s4_pos1 ("The site appears uniform") was thus picked because it was favored in the language center evaluation for being a closer antonym to VisAWI_s4 ("The site appears patchy") than VisAWI_s4_pos2 ("The site appears consistent"). In contrast, item *VisAWI_d1_pos1* ("The design is interesting") was chosen as an alternative for *VisAWI_d1* ("The design is uninteresting") based not only on a preference in the language center evaluation but also on a higher correlation between the two, in contrast to item *VisAWI_d1_pos2* ("The design is exciting"). All 18 items for the VisAWI-pos can be seen in Table 4. The complete rationales for item selection are detailed in the analysis script on OSF.

4.2.3. Initial comparison of VisAWI and VisAWI-pos

After settling on 18 items for the VisAWI-pos, these items were compared to the original VisAWI items. Here, the analysis focused on descriptive statistics, internal consistency as an indicator of reliability (coefficient α, Cronbach, 1951), the ability to differentiate between the two website conditions, and correlations.

Descriptive statistics for the VisAWI-pos were comparable to those of the original VisAWI, both overall and within the two conditions (see Table 5). Concerning reliability, results for both the original VisAWI and the VisAWI-pos were excellent (George & Mallery, 2019) and of comparable magnitude. Results for the individual sub-scales were comparably good for the two versions, with slightly higher values for the VisAWI-pos (see Table 5). Furthermore, the analysis considered if the two versions of the scale could distinguish between the stimuli websites, using Welch's two-sample t-tests for unequal variances. Results showed that the aesthetic variant scored significantly higher on the original VisAWI total score than the unaesthetic variant, $t(30.56) = 7.20, \mathrm{p} < 0.0001, d = 2.32$. For the overall score of the VisAWI-pos, the difference was also significant, with the aesthetic variant scoring significantly

Table 5. Internal consistency (α) and descriptive statistics from study one for the original VisAWI and the VisAWI-pos (overall and per condition, aesthetic vs. unaesthetic).

		Overall ($N = 41$)				Aesthetic ($n = 22$)				Unaesthetic ($n = 19$)			
	α	Mean	SD	min	max	Mean	SD	min	max	Mean	SD	min	max
VisAWI	0.97	4.53	1.70	1.31	6.95	5.72	0.88	3.41	6.95	3.15	1.32	1.31	6.04
VisAWI-pos	0.98	4.50	1.85	1.25	6.89	5.81	0.88	3.22	6.89	2.99	1.48	1.25	6.00
Simplicity	0.88	4.93	1.57	2.00	7.00	5.89	0.94	3.80	7.00	3.81	1.42	2.00	6.20
Simplicity (pos)	0.92	4.86	1.72	1.00	7.00	6.02	0.80	3.80	7.00	3.52	1.50	1.00	6.00
Diversity	0.87	4.31	1.66	1.00	7.00	5.29	1.21	2.60	7.00	3.18	1.39	1.00	5.60
Diversity (pos)	0.94	4.38	1.76	1.00	7.00	5.44	1.11	2.60	7.00	3.15	1.57	1.00	6.00
Colorfulness	0.94	4.30	2.16	1.00	7.00	5.93	0.92	3.50	7.00	2.41	1.52	1.00	6.50
Colorfulness (pos)	0.97	4.16	2.22	1.00	7.00	5.78	1.05	3.00	7.00	2.28	1.65	1.00	6.00
Craftsmanship	0.85	4.58	1.76	1.25	7.00	5.77	0.92	3.75	7.00	3.20	1.47	1.25	6.25
Craftsmanship (pos)	0.94	4.62	1.97	1.00	7.00	5.99	0.95	3.50	7.00	3.04	1.63	1.00	6.00

higher than the unaesthetic variant, $t(28.43) = 7.23, p < 0.0001, d = 2.35$. Concerning the subscales, all differences between the conditions were significant for both scale versions (effect sizes d between 1.63 and 2.85). Finally, correlations between the two versions of the scale were looked at. The overall scores for the VisAWI and the VisAWI-pos correlated almost perfectly ($r = 0.99$). Correlations among the individual subscales of the two versions were also very high: $r = 0.96$ for simplicity, $r = 0.95$ for diversity, $r = 0.97$ for colorfulness, and $r = 0.96$ for craftsmanship.

4.3. Discussion

After the first study, it was concluded that there now was an initial set of positive alternative items for the VisAWI comparable to the original negatively formulated items, producing almost identical results. This new alternative version, the *VisAWI-pos*, consists of the ten original positively formulated VisAWI items and eight newly developed positive alternatives. As a next step, the psychometric quality of the VisAWI-pos was to be investigated with a larger sample, comparing it to the VisAWI and VisAWI-S. Furthermore, while using manipulated websites allowed the examination of how well the VisAWI and the VisAWI-pos could differentiate between two versions of a website, the manipulation of the website arguably caused responses limited to the bottom and top of the scale instead of distributions across the entire range of answering options. Therefore, further research on the VisAWI-pos was necessary.

5. Study two: Psychometric evaluation

As a next step, the aim was to investigate the psychometric quality of the VisAWI-pos with a sufficiently large sample and compare its quality to the original VisAWI and VisAWI-S. In addition, it was investigated how the different VisAWI versions would work with existing websites, as opposed to sites manipulated for the context of an experiment. Therefore, the second study addressed the following two research questions: How does the VisAWI perform concerning common indicators of psychometric quality? And how does the psychometric quality of the VisAWI-pos compare to that of the original VisAWI and VisAWI-S? To investigate these questions, a second study with a considerably larger sample and a set of existing websites used as stimuli was conducted. This study was pre-registered on OSF (https://osf.io/u5w7t/).

5.1. Methods

A between-subjects online study was conducted. Participants interacted with a website randomly drawn from a pool of 12 existing websites and responded to several questions related to the website's content. After the interaction, participants were presented with all items of the VisAWI and VisAWI-pos and a selection of survey scales to assess the VisAWI versions' convergent and divergent validity.

5.1.1. Stimuli

As stimuli, a set of 12 websites from Australia were prepared. Australian sites were selected to minimize participants' familiarity with the site while still being able to understand the site's language, given that participants from the United States of America (USA) and the United Kingdom (UK) were recruited. Participants were also asked if they already knew the site they visited to check for possible differences between those familiar with the site and those unfamiliar. The 12 sites covered six content areas (arts and entertainment, law & government, news & media publishers, science & education, food & drink, and lifestyle). Content areas were taken from Similarweb.com and selected to cover various types of websites. For each content area, one popular and one unpopular website was chosen based on rankings from Similarweb.com. Popular sites were among the top 100 in Australia, while unpopular sites had lower ranks, ranging from 302 to 73,194. Table 6 contains all 12 websites used, including their content area and rank.

For each site, participants had to answer two content-based questions specifically developed for the individual websites. The first question generally asked what the website was about (eg, news, fashion, weather). The second question consisted of a site-specific task to be completed, with four answer options, of which only one was correct (eg, "You want to buy a gift card to give to a friend. What amounts are there?"). The tasks were designed to require a minimum of one to four clicks to be completed, and answers could not be found by just looking at the site's landing page. This was done to have a comparable effort to complete each task across conditions. All tasks are provided in the

Table 6. Websites used as stimuli in study two.

Website name	Link	Content area	Rank
Stan.com	https://www.stan.com.au/	Arts & entertainment	92
NSW Government	https://www.nsw.gov.au/	Law & government	25
ABC News	https://www.abc.net.au/	News & media publishers	17
Australian Government Bureau of Meteorology	http://www.bom.gov.au/	Science & education	22
Woolworths	https://www.woolworths.com.au/	Food & drink	23
The Iconic	https://www.theiconic.com.au/	Lifestyle	63
AussieTheater.com	https://www.aussietheatre.com.au/	Arts & entertainment	20,584
Government of Western Australia	https://www.wa.gov.au/	Law & government	1,011
Great Lakes Advocate	https://www.greatlakesadvocate.com.au	News & media publishers	73,194
Willy Weather	https://www.willyweather.com.au/	Science & education	302
IGA Supermarkets	https://www.iga.com.au/	Food & drink	1,179
Jeanswest Australia	https://www.jeanswest.com.au/	Lifestyle	2,691

Website popularity ranks and content areas were retrieved on 22.09.2022 from Similarweb.com.

supplementary materials on OSF. Item difficulty for the questions ranged from 62.96% correct answers to 100.00%, with a mean of 93.03% correct answers ($SD = 10.06\%$). Across all 12 websites, 98.24% of participants stated they were unfamiliar with the site (range for the individual sites from 93.67% to 100.00% not familiar). Given that most participants were unfamiliar with the stimuli sites, data for all participants, familiar and unfamiliar, were analyzed together.

5.1.2. Participants

For the experiment, 1003 participants from the UK and USA were recruited on Prolific and paid £1.50. Participants were restricted to having no color blindness or visual impairment based on the information given via Prolific. Furthermore, they were only allowed to participate if they did not partake in the prior study. During data cleaning, 32 observations were removed because they failed an instructed response item and another three due to a seriousness check, based on recommendations for assuring data quality in online surveys from Brühlmann et al. (2020). In addition, one participant was excluded for reporting an unrealistic age (four) and one participant for indicating their current country of residence outside the USA or UK, going against the recruitment criteria. Data cleaning resulted in a final sample of 966 respondents (550 women, 407 men, seven non-binary, one not specified, one self-described; mean age 41.11 years, $SD = 13.49$, $min = 18$, $max = 80$; 920 from the UK and 46 from the USA).

5.1.3. Measures

Participants rated the stimuli website by responding to all VisAWI and VisAWI-pos items. In addition, several survey scales were used to assess the VisAWI versions' convergent and divergent validity. A comparable pattern of correlations with these scales was expected for all three VisAWI versions (original, positive, and short). The selection of scales was based on past work validating the original VisAWI, or versions of it (Abbas et al., 2022; Moshagen & Thielsch, 2010, 2013).

5.1.3.1. Visual aesthetics scale. Data on visual aesthetics was collected using the scale by Lavie and Tractinsky (2004) as a convergent measure. The scale consists of ten items, measuring two separate aesthetic constructs, classic and expressive aesthetics, with five items each. Responses to the items were collected on a seven-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). The scale exhibited very high internal consistency for the classical aesthetics items ($\alpha = 0.90$, 95% CI[0.89, 0.91], $\omega = 0.90$, 95% CI[0.89, 0.91]) and expressive aesthetics ($\alpha = 0.89$, 95% CI[0.87, 0.90], $\omega = 0.89$, 95% CI[0.88, 0.90]). It was decided to use this scale because it is among the only aesthetics scales frequently used in HCI research besides the VisAWI (Thielsch et al., 2019), and it was also used for the assessment of convergent validity during the initial development of the VisAWI. Given that both the VisAWI and this scale are supposed to measure visual aesthetics, high correlations between the two were expected. Furthermore, based on past research (Moshagen & Thielsch, 2010), higher correlations between the overall VisAWI and classic aesthetics than with expressive aesthetics were expected. For the individual facets of the VisAWI, diversity was expected to correlate higher with expressive aesthetics. In contrast, the other three facets were predicted to correlate higher with the classic aesthetics score.

5.1.3.2. User Experience Questionnaire. The User Experience Questionnaire (UEQ, Laugwitz et al., 2008) was used to measure several UX-related constructs, both convergent and divergent to the VisAWI. The UEQ is a semantic differential scale comprising 26 adjective pairs, to which responses were collected on a seven-point scale. After data collection, the values were re-coded to range from −3 to +3 in line with recommendations by the original authors. Items are grouped across six dimensions, with six items for the overall attractiveness of a product and four items each for the three goal-oriented pragmatic quality dimensions (ie, perspicuity, efficiency, dependability) and the two non-goal-directed hedonic quality dimensions (ie, stimulation, novelty). Internal consistency was good for the overall UEQ ($\alpha = 0.96$, 95% CI[0.95, 0.96], $\omega = 0.96$, 95% CI[0.95, 0.96]), and for the individual subscales (values for α and ω from 0.71 to 0.96, see OSF for details). It was decided to work with the UEQ because it promises to cover a broad selection of constructs relevant to users' experiences. While Moshagen and Thielsch (2010) worked with the AttrakDiff (Hassenzahl, 2004) when developing the original VisAWI, the present study worked with the more recent UEQ, which builds upon the same theoretical model as the AttrakDiff.

Furthermore, the UEQ was also used recently by Abbas et al. (2022) while validating the AR-VisAWI. Following results from prior research (Abbas et al., 2022; Moshagen & Thielsch, 2010, 2013), the pragmatic quality constructs of the UEQ were expected to correlate to a lesser extent with the VisAWI ratings than the rating of attractiveness. The only exception was the simplicity facet, for which comparable correlations with attractiveness and pragmatic quality were expected, based on Moshagen and Thielsch (2010). In addition, higher correlations between the VisAWI and the hedonic quality constructs were expected compared to pragmatic quality.

5.1.3.3. SUS – positive version. The positive version of the SUS (Sauro & Lewis, 2011) was used as an indicator of participants' perceived usability. Perceived usability was measured with this particular scale to avoid reverse-coded items. The positive SUS consists of ten items, to which answers were collected using a five-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Following Brooke (1996), responses to the ten items were transformed into scores ranging from 0 to 100 before interpretation. The scale exhibited very high internal consistency ($\alpha = 0.92$, 95% CI[0.91, 0.93], $\omega = 0.92$, 95% CI[0.91, 0.93]). Usability was considered a divergent construct in past psychometric evaluation attempts for the VisAWI (Abbas et al., 2022; Moshagen & Thielsch, 2010, 2013). Based on this, a positive correlation between the SUS score and the VisAWI ratings was expected, although to a lesser extent than with visual aesthetics, thus taking the SUS as an indicator of the VisAWI's divergent validity. In addition, concerning the sub-scale simplicity, higher correlations with the SUS than for the other sub-scales of the VisAWI were expected based on past VisAWI validations.

5.1.3.4. Web-CLIC-S. To judge the stimuli websites' content, the Web-CLIC-S (Thielsch & Hirschfeld, 2021) was used, a short version of the Web-CLIC scale. The Web-CLIC-S consists of four items, one for each of the Web-CLIC's content areas: clarity, likability, informativeness, and credibility. Items are rated on a seven-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree). Mean values across these four items were formed to measure participants' subjective content perception of the stimuli websites. Internal consistency for the scale was good ($\alpha = 0.84$, 95% CI[0.82, 0.86], $\omega = 0.85$, 95% CI[0.83, 0.87]). The Web-CLIC-S was used because "quality of content" was also considered during the original VisAWI's development, although assessed with another set of items. The Web-CLIC-S was chosen over this other set of items due to its briefness, reducing participant burden, and because it is relatively new, which makes it more suitable for current websites. In addition, the VisAWI was also used as a divergent measure during the development of the Web-CLIC-S (Thielsch & Hirschfeld, 2021). Based on past findings (Thielsch & Hirschfeld, 2021), lower correlations between the VisAWI ratings and the Web-CLIC-S than for visual aesthetics were expected.

5.1.3.5. Net Promoter Score. Participants also responded to the single-item Net Promoter Score (NPS, Reichheld, 2003). The NPS asks users a single question about the likelihood of recommending the website to a friend or colleague, taken as an indicator of customer loyalty. Responses are recorded on an eleven-point Likert-type rating scale ranging from 0 to 10. The raw NPS rating can then be transformed into an NPS score ranging from -100 to $+100$. Responses to the single-item NPS were collected as an indicator of divergent validity for the VisAWI and VisAWI-pos. A positive correlation between the VisAWI scores and the NPS was expected, although lower than between the VisAWI and VisAWI-pos scores or than with the convergent measures.

5.1.4. Procedure

Participants provided informed consent on the first page of the survey. Next, they were given instructions for the task to be completed. Participants were then assigned to one of the 12 websites and asked to explore the website, responding to the questions related to the website content. Afterward, they filled out the 18 items of the original VisAWI and the VisAWI-pos' eight positive alternatives, shown in randomized order. On the following pages of the survey, participants responded to the convergent and divergent survey scales in randomized order. Next, participants provided demographic information (age, gender, country of residence). Lastly, they had the opportunity to give feedback and then were redirected to Prolific for payment. Two instructed response items (Curran, 2016) embedded among the survey scales and a single item for self-reported data quality (Meade & Craig, 2012) at the end of the survey were used to ensure high response quality, following recommendations by Brühlmann et al. (2020). Figure 2 gives an overview of the study procedure. Participants took an average of 10.18 minutes to complete the survey ($SD = 4.82$, $min = 3.87$, $max = 49.42$). Based on past findings (Schriesheim & Hill, 1981), the negatively worded items were expected to have no relevant effect on participants' responses to the positively worded items of the scale. The study thus had participants fill in all of the original items of the VisAWI, both positively and negatively worded, alongside the positive alternatives, resulting in a larger sample, especially crucial for confirmatory factor analysis (CFA) (Kline, 2016).

5.1.5. Pre-Test

Before pre-registering the study, the procedure, stimuli, and tasks were tested with a small-sample pre-study ($N = 36$, three per stimuli website). The main goal of this pre-test was to examine if most participants would be able to complete the tasks and if there were any major issues in the study procedure. The participants answered most tasks correctly, with a mean of 88.89% correct answers ($SD = 25.38\%$). However, two tasks were adjusted for the full-sample study, which most pre-study participants in the respective conditions did not complete successfully.[2]

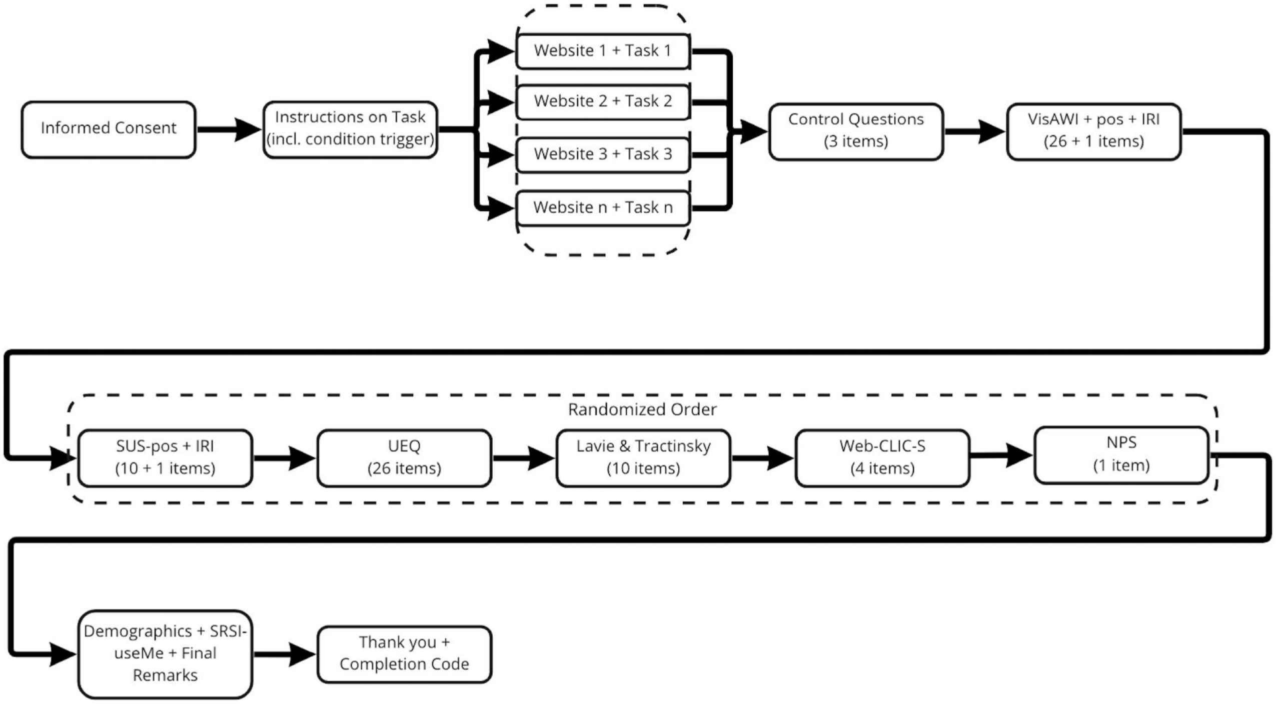

Informed Consent
Instructions on Task (incl. condition trigger)
Website 1 + Task 1
Website 2 + Task 2
Website 3 + Task 3
Website n + Task n
Control Questions (3 items)
VisAWI + pos + IRI (26 + 1 items)
Randomized Order
SUS-pos + IRI (10 + 1 items)
UEQ (26 items)
Lavie & Tractinsky (10 items)
Web-CLIC-S (4 items)
NPS (1 item)
Demographics + SRSI-useMe + Final Remarks
Thank you + Completion Code

Figure 2. Flowchart for the procedure of study two.

Table 7. Coefficients α and ω from study two for the VisAWI, the VisAWI-pos, and the VisAWI-S (including 95% CIs).

	coefficient α			coefficient ω		
Sub-scale	VisAWI	VisAWI-pos	VisAWI-S	VisAWI	VisAWI-pos	VisAWI-S
Overall	0.96 [0.95, 0.96]	0.96 [0.96, 0.96]	0.87 [0.86, 0.88]	0.95 [0.95, 0.96]	0.96 [0.96, 0.96]	0.87 [0.85, 0.89]
Simplicity	0.87 [0.86, 0.89]	0.90 [0.89, 0.91]	–	0.87 [0.86, 0.89]	0.90 [0.89, 0.91]	–
Diversity	0.89 [0.88, 0.91]	0.90 [0.89, 0.91]	–	0.90 [0.89, 0.91]	0.90 [0.89, 0.91]	–
Colorfulness	0.88 [0.86, 0.89]	0.90 [0.89, 0.91]	–	0.88 [0.86, 0.89]	0.91 [0.90, 0.92]	–
Craftsmanship	0.86 [0.85, 0.88]	0.89 [0.87, 0.90]	–	0.86 [0.85, 0.88]	0.89 [0.88, 0.90]	–

5.2. Results

The following section contains different forms of psychometric quality investigation for the three VisAWI versions: original, positive, and short. The complete analysis can be found on OSF (https://osf.io/x596a). All results were obtained using the statistical software R (version 4.2.2).

5.2.1. Item analysis

Concerning item analysis, descriptive statistics, item difficulty and variance, discriminatory power, and inter-item correlations for all 26 VisAWI items (18 original and eight positive alternatives) were considered. In summary, item analysis showed that while values sometimes differed between the original VisAWI items and their positive counterparts, no items exhibited values considered to be overtly problematic. The analysis thus continued without flagging or removing any items.

5.2.2. Reliability

Following recommendations by Dunn et al. (2014), both coefficients α (Cronbach, 1951) and ω (McDonald, 1999) were calculated as indicators of reliability for all three VisAWI versions (original, positive, and short), including the sub-scales. Values for both coefficients (including 95% CIs) are presented in Table 7. All values were above 0.80, with multiple values above 0.90, indicating very high internal consistency. Furthermore, all values for the VisAWI-pos were equal to or slightly higher than for the original VisAWI and the VisAWI-S.

5.2.3. Confirmatory factor analyses

One CFA each for the VisAWI and VisAWI-pos was performed to investigate the fit to the original model (ie, four factors with one higher-order factor). For the VisAWI-S, a single-factor model was used. Because the Henze-Zirkler test (Henze & Zirkler, 1990) and Mardia's test (Mardia, 1970) indicated a violation of the multivariate normality assumption for all VisAWI versions, a robust maximum likelihood estimator was used. The following criteria were considered as an indication of good model fit: Low χ^2 value and $p > 0.05$ for the Chi-squared test, $RMSEA < 0.06$, $SRMR \leq 0.08$ and $0.95 \leq CFI \leq 1$ (Hu & Bentler, 1999).[3] All goodness-of-fit statistics from the CFA results are presented in Table 8.

Starting with the original VisAWI, only the *SRMR* indicated a good fit, while all other CFA fit measures suggested otherwise. Figure A1 in the Appendix shows the CFA model used for the original VisAWI, including all loadings. Turning to the VisAWI-pos, CFA results suggested an inadequate fit based on the χ^2-test and the *RMSEA*, while the

SRMR and *CFI* favored the model. Nevertheless, fit indices for the VisAWI-pos were identical in the case of *SRMR* and otherwise better than those calculated for the original VisAWI (larger *CFI*, smaller χ^2, and smaller *RMSEA*). In addition to the sub-optimal fit, the model exhibited a negative residual variance (–0.003) for the craftsmanship facet (ie, a Heywood case, Heywood, 1931). Comparable Heywood cases on the craftsmanship facet were also reported for the AR-VisAWI (Abbas et al., 2022). The model used for the VisAWI-pos is shown in the Appendix in Figure A2. Finally, the VisAWI-S was analyzed using a single-factor model (see Figure A3 in the Appendix for the complete model). The CFA suggested a good overall fit. Only the *RMSEA* and the significant χ^2-test indicated an inadequate fit, while all other indices were in support of the model.

To conclude, out of all CFAs, the VisAWI-S showed the most favorable results, closely followed by the VisAWI-pos. While the *RMSEA* for the VisAWI-pos was slightly above the ideal cutoff of 0.06, *RMSEA* values < 0.08 can still be considered a reasonable error of approximation (MacCallum et al., 1996). Moreover, the χ^2 test is known to be influenced by larger sample sizes (> 200) and deviations from multivariate normality, both present in the current case (Whittaker & Schumacker, 2022). Therefore, the significant χ^2 test has reduced informativeness for all three VisAWI versions. Thus, the overall fit of the VisAWI-pos was better than for the original VisAWI, where only the *SRMR* indicated a good model fit. Nevertheless, the sub-optimal fit indices for the VisAWI and the Heywood case for the craftsmanship factor of the VisAWI-pos warranted further investigation. Consequently, exploratory factor analyses (EFAs) were conducted to search for alternative models for the VisAWI and VisAWI-pos.

Table 8. Fit Indices for CFA models of the VisAWI versions in study two.

Model	χ^2	df	p-value χ^2	RMSEA	SRMR	CFI
VisAWI	792.30	131	<0.001	.083	.047	.933
VisAWI-pos	576.10	131	<0.001	.067	.047	.960
VisAWI-S	13.33	2	<0.01	.091	.017	.991

Robust values are reported wherever possible.

5.2.4. Exploratory factor analyses

Multiple EFAs were run to identify alternative factor structures for the VisAWI and VisAWI-pos items to see if there are ways to improve the model fit while keeping the original theoretical model in mind. Following Moshagen and Thielsch (2010), the *oblimin* rotation method (an oblique rotation) and principal axis factoring as the extraction method were chosen. In line with recommendations by Howard (2016), the $0.40 - 0.30 - 0.20$ rule was used to interpret factor loadings, which states that an item should load at least 0.40 on a primary factor with no secondary loadings > 0.30 and differences of at least 0.20 between the primary loading and any secondary loadings. Concerning the interpretation of communality, values < 0.50 were considered as problematic (Hair et al., 2010). For both the VisAWI and VisAWI-pos data, prerequisites for EFA were given: Bartlett's test for sphericity was significant (original VisAWI: $\chi^2(153) = 13,124.95$, $p < 0.001$; VisAWI-pos: $\chi^2(153) = 14,163.67$, $p < 0.001$) and the Kaiser-Meyer-Olkin Measures of Sampling Adequacy were great for the VisAWI and VisAWI-pos (all > 0.90, Howard, 2016). Following the theoretical model of the VisAWI, the analysis focused on four-factor solutions for both the VisAWI and VisAWI-pos. Additional solutions based on parallel analyses and scree-plots were also considered, for which interested readers are referred to OSF.

5.2.4.1. VisAWI. The analysis first started with the original VisAWI. The four-factor solution was able to explain 66% of cumulative variance but showed problematic loading patterns for six items. Table 9 contains all factor loadings > 0.20 and communalities for the four-factor EFA. While three out of four craftsmanship items loaded adequately on one factor (> 0.40) without any high cross-loadings, they did not load onto a separate factor but rather on the same

Table 9. Factor loadings > 0.20 and communalities from study two of all 18 original VisAWI items with the four-factor model.

Item name	Item	PA2	PA1	PA3	PA4	h2
VisAWI_s1	The layout appears too dense. (R)	0.40			0.24	0.38
VisAWI_s2	The layout is easy to grasp.	0.88				0.64
VisAWI_s3	The layout appears well structured.	0.87				0.78
VisAWI_s4	The site appears patchy. (R)	0.46	0.30		0.28	0.62
VisAWI_s5	Everything goes together on this site.	0.66				0.69
VisAWI_d1	The design is uninteresting. (R)		0.78			0.74
VisAWI_d2	The layout is inventive.		0.66			0.58
VisAWI_d3	The design appears uninspired. (R)		0.88			0.74
VisAWI_d4	The layout appears dynamic.		0.52			0.57
VisAWI_d5	The layout is pleasantly varied	0.41	0.26	.29	−0.21	0.66
VisAWI_col1	The color composition is attractive.			.86		0.84
VisAWI_col2	The choice of colors is botched. (R)			.47	0.46	0.68
VisAWI_col3	The colors do not match. (R)			.40	0.44	0.60
VisAWI_col4	The colors are appealing.			.85		0.81
VisAWI_craf1	The layout appears professionally designed.	0.46	0.30			0.65
VisAWI_craf2	The layout is not up-to-date. (R)		0.64			0.57
VisAWI_craf3	The site is designed with care.	0.58	0.25			0.72
VisAWI_craf4	The design of the site lacks a concept. (R)	0.24	0.55			0.60

Problematic items are marked in bold and reverse-coded items with (R).
PA1–PA4 = factor loadings; h2 = communality.

Table 10. Pearson's product-moment correlations from study two between the VisAWI versions and with the convergent and divergent scales (including 95% CIs).

	VisAWI	VisAWI-pos	VisAWI-S
VisAWI-pos	0.97 [0.97, 0.98]		
VisAWI-S	0.94 [0.94, 0.95]	0.96 [0.95, 0.96]	
Classic aesthetics	0.83 [0.81, 0.85]	0.85 [0.83, 0.87]	0.81 [0.79, 0.83]
Expressive aesthetics	0.73 [0.69, 0.75]	0.74 [0.71, 0.77]	0.70 [0.67, 0.73]
Attractiveness (UEQ)	0.84 [0.82, 0.86]	0.85 [0.83, 0.87]	0.82 [0.80, 0.84]
Hedonic quality (UEQ)	0.71 [0.67, 0.74]	0.72 [0.69, 0.75]	0.67 [0.63, 0.70]
Pragmatic quality (UEQ)	0.75 [0.72, 0.78]	0.75 [0.72, 0.77]	0.71 [0.68, 0.74]
Usability (SUS score)	0.68 [0.65, 0.71]	0.70 [0.67, 0.73]	0.64 [0.61, 0.68]
Web-CLIC-S	0.76 [0.73, 0.79]	0.78 [0.76, 0.81]	0.74 [0.71, 0.77]
NPS (raw values)	0.70 [0.66, 0.73]	0.71 [0.67, 0.74]	0.67 [0.63, 0.70]

All correlations were significant at $p < 0.0001$.

factors as the simplicity or diversity items. Furthermore, it was notable that out of six ill-fitting items, four were negatively formulated, the fourth factor explained little variance (5%), and the negative item VisAWI_s1 had a low communality ($h2 = 0.38$).

From the EFA, it was deduced that the negative items were a probable reason for the sub-optimal fit in the CFA for the original VisAWI. In addition, the items for craftsmanship did not form a clear factor on their own, pointing to more fundamental issues with the craftsmanship facet. It was thus concluded that the items for simplicity and diversity mostly formed distinct factors. Regarding colorfulness, item tone caused items to load onto separate factors, while the craftsmanship items did not load onto a unique factor. Overall, a four-factor solution appeared most suitable for the scale. Nevertheless, results also suggested that future work would have to investigate if the VisAWI benefits from certain adjustments, such as item removal or refinement and reinvestigation of the scale's theoretical structure, for an improved fit.

5.2.4.2. VisAWI-pos.

Next, a four-factor solution for the VisAWI-pos was investigated, following the original VisAWI model. The four-factor solution was able to explain 69% of the cumulative explained variance. Communalities for all items were > 0.50, except for item VisAWI_s4_pos, which was slightly lower ($h2 = 0.48$). Table 4 contains all factor loadings > 0.20 and all communalities for the four-factor EFA. Items for simplicity, diversity, and colorfulness loaded well onto three separate factors. However, there were issues with the loadings for three out of four craftsmanship items. In addition, while the item VisAWI_craf4_pos exhibited a high primary loading without any relevant secondary loadings, it loaded onto the same factor as the simplicity items instead of a distinct factor. Furthermore, the fourth factor explained just 7% of the variance. From these results, it was concluded that a four-factor solution was reasonable for the VisAWI-pos despite issues with the craftsmanship items likely also responsible for the Heywood case in the CFA. These issues concerning the craftsmanship facet will be revisited in the discussion section.

5.2.5. Convergent and divergent validity

For the assessment of convergent and divergent validity, two approaches were used: First, Pearson's product-moment correlations were calculated. Second, CFAs were performed with all scales included, calculating error-free correlations among variables (Eid et al., 2017). Correlation patterns with the other scales were comparable for the original VisAWI, the VisAWI-S, and the VisAWI-pos and thus are described here by just talking about the VisAWI (see Table 10 for exact values). Correlations for the VisAWI sub-scales are provided in the Appendix (see Table A1). Detailed values, including correlations among the facets of the VisAWI versions and with the UEQ sub-scales, can be found in the supplementary materials on OSF.

Overall, most of the correlations were as expected, thus favoring the VisAWI's convergent and divergent validity for all three versions. Concerning convergent validity, the data exhibited the expected high correlations between the VisAWI score and classic aesthetics, as well as expressive aesthetics, with higher correlations for classic aesthetics. As expected, the VisAWI sub-scales correlated higher with classic aesthetics, except for diversity, which correlated higher with expressive aesthetics. Correlations with the SUS were also as predicted for all three VisAWI versions. Compared to the other aesthetics scale, the correlations of the VisAWI with the SUS were lower than correlations of the VisAWI with classic aesthetics but just slightly lower than for expressive aesthetics. Regarding the UEQ, attractiveness correlated higher than pragmatic quality, which was in line with expectations. Interestingly, hedonic quality showed lower correlations with the VisAWI compared to pragmatic quality. Regarding the Web-CLIC-S, all correlations were comparably high, both for the overall VisAWI scores and the individual sub-scales. As expected, the overall VisAWI scores correlated to a lesser extent with the Web-CLIC-S than with classic aesthetics, although correlations with expressive aesthetics were of a comparable magnitude. Correlations with the NPS were comparable across VisAWI versions and sub-scales and mostly lower than with the convergent constructs.

With the large CFA, comparable results to Pearson's correlations were found, which provided further evidence for the scales' convergent and divergent validity. Details of these measurement-error-free correlations can be found in the Appendix (see Table A2). Overall, these results supported the convergent and divergent validity of all three VisAWI versions.

Table 11. Descriptive statistics from study two for the three VisAWI versions (original, positive, short) by website condition (popular = 1–6, unpopular = 7–12), including rankings (based on mean).

		Original VisAWI					VisAWI-pos					VisAWI-S				
website	*n*	mean	SD	Min	max	rank	mean	SD	min	max	rank	mean	SD	min	max	rank
1	79	5.30	1.02	2.49	7.00	2	5.23	1.01	2.41	7.00	2	5.30	1.07	2.50	7.00	2
2	83	5.13	0.88	2.90	6.64	3	5.04	0.86	2.72	6.54	3	5.18	1.00	2.00	6.75	3
3	81	4.49	1.06	2.46	6.74	9	4.46	1.02	2.15	6.65	10	4.58	1.16	1.75	7.00	8
4	81	4.91	1.18	1.10	7.00	5	4.90	1.14	1.31	7.00	5	4.99	1.20	1.00	7.00	5
5	79	4.79	1.18	1.32	6.89	7	4.76	1.17	1.10	6.78	6	4.76	1.22	1.00	7.00	6
6	82	5.50	0.69	4.00	6.72	1	5.42	0.63	4.00	6.58	1	5.58	0.74	3.50	6.75	1
7	80	4.49	1.23	1.79	7.00	8	4.47	1.24	1.56	7.00	8	4.57	1.33	1.50	7.00	9
8	80	4.46	0.92	2.52	7.00	10	4.47	0.89	2.19	7.00	9	4.46	1.02	2.00	7.00	10
9	78	4.30	1.04	1.74	7.00	11	4.30	1.06	1.35	7.00	11	4.36	1.19	1.25	7.00	11
10	80	4.18	1.19	1.89	7.00	12	4.14	1.21	1.39	7.00	12	4.03	1.37	1.00	7.00	12
11	83	4.79	1.06	2.11	7.00	6	4.76	1.05	1.99	7.00	7	4.75	1.17	1.50	7.00	7
12	80	4.96	0.98	1.89	6.46	4	4.93	0.93	1.68	6.47	4	5.13	1.03	1.25	7.00	4

5.2.6. Differentiation between stimuli websites

Finally, to investigate the VisAWI versions' ability to distinguish between different websites, the analysis looked for statistical differences in VisAWI scores between the twelve stimuli websites. Descriptive statistics for ratings of the VisAWI, VisAWI-pos, and VisAWI-S, sorted by condition, can be found in Table 11.

The researchers started by looking for significant differences in VisAWI ratings between the twelve websites. Because results were comparable for all three VisAWI versions, they are only reported here for the VisAWI-pos while interested readers are referred to OSF for details on the VisAWI and VisAWI-S. Concerning the VisAWI-pos overall score, an ANOVA found significant differences ($F(11, 954) = 11.32,\ p < 0.001,\ \eta^2 = 0.12$). In addition, a Kruskal–Wallis rank sum test was performed because equal variances and normal distribution were not given, which was likewise significant ($\chi^2(11) = 119.46,\ p < 0.001$). Tukey's HSD Test found multiple significant individual differences between the websites, provided in detail on OSF. Ratings of the websites for the VisAWI-pos sub-scales also differed significantly ($\eta^2_{\min} = 0.07, \eta^2_{\max} = 0.12$).

Furthermore, it was considered how the 12 stimuli websites are ranked based on the mean ratings (from highest to lowest mean) and if those rankings are comparable across the scale version (eg, if the highest-rated site using the VisAWI was also rated highest with the VisAWI-pos and VisAWI-S). The rankings based on means are reported in Table 11. The five highest-ranked sites were the same across all three scale versions. Similarly, the two lowest-ranked sites were also identical. Concerning the rankings in between, results were comparable, with only slight changes in rank orders between the three scales. Thus, rankings based on mean overall ratings were mostly identical across the three scale versions.

Finally, the VisAWI cutoff of $\bar{x} \geq 4.50$ for aesthetic websites suggested by Hirschfeld and Thielsch (2015) was used to see if the three different scale versions would deliver comparable results regarding this cutoff. For the original VisAWI, seven out of 12 websites were above the cutoff, while three (no. three, seven, and eight) were just below it ($4.46 - 4.49$). For the VisAWI-pos, the same seven websites were above the cutoff, while the same three websites as with the VisAWI had a slightly lower rating ($4.46 - 4.47$). Concerning the VisAWI-S ratings, the same seven websites again had mean ratings of $\bar{x} \geq 4.50$. Two of the three websites rated just below the cutoff using the full versions (no. three and seven) also had mean ratings above the cutoff, while the third website (no. eight) again fell slightly below it ($\bar{x} = 4.46$). Thus, mean ratings for all three VisAWI versions delivered comparable results concerning the VisAWI cutoff for aesthetic websites.

From these results, it was concluded that all three versions of the VisAWI could differentiate between the stimuli websites used in the study while delivering very similar results, providing evidence for the scales' criterion-related validity.

6. General discussion

The present work developed a version of the VisAWI free of reverse-coded items. While reverse-coded items are intended to counteract undesirable response behavior, results from past research have shown these items to do more harm than good (Dalal & Carter, 2014; Sauro & Lewis, 2011). Thus, the present work set out to improve the quality of the VisAWI by developing a version with only positive items called the VisAWI-pos, which was developed and then validated throughout two studies. As a first step, an initial set of positively formulated alternatives for the eight reverse-coded items of the VisAWI were created by the researchers and reviewed by a language expert. A first study using two aesthetically manipulated websites showed that most candidate items were viable alternatives, performing comparably to the original VisAWI items. Thus, eight positive alternatives, one for each reverse-coded VisAWI item, were chosen based on the results from the first study and the language evaluation. Finally, a second study using existing websites was conducted to evaluate the psychometric quality of the VisAWI-pos, comparing it to the VisAWI and VisAWI-S. Findings from this second study showed that the VisAWI-pos delivered comparable results to the original VisAWI versions while performing better in terms of reliability and validity. As part of this second study, the psychometric quality of the English version of the VisAWI and the VisAWI-S was also investigated for the first time, showing that the two scales are of comparable quality to versions of the scale published in other languages (ie, German, Arabic, and Farsi).

6.1. VisAWI-pos – the better alternative

Two studies demonstrated that the VisAWI-pos is as good or better than other versions of the scale, with the added benefit of avoiding issues associated with reverse-coded items. Model fit indices from CFA mostly favored the VisAWI-pos, with $SRMR \leq 0.08$, $CFI > 0.95$, and only $RMSEA$ slightly above the cutoff of < 0.06 suggested by Hu and Bentler (1999), but below 0.08 which is indicative of a reasonable error of approximation (MacCallum et al., 1996). Moreover, the significant χ^2 test for all three VisAWI versions should not be a cause for concern, as the test is sensitive to deviations from multivariate normality and tends to be significant for larger samples with more than 200 participants (Whittaker & Schumacker, 2022). Concerning the χ^2 value, only the VisAWI-pos showed a ratio of $\chi^2/df < 5$, sometimes considered indicative of a good model fit (Wheaton et al., 1977). Furthermore, the VisAWI-pos' model fit was comparable to what was reported for the original German version of the VisAWI (Moshagen & Thielsch, 2010),[4] the Arabic version (Abbas et al., 2022),[5] and the Farsi version (Saremi et al., 2023).[6] Thus, the findings showed a good fit of the VisAWI-pos data to the theoretical model suggested by Moshagen and Thielsch (2010), considering common model fit indicators. Finally, the VisAWI-pos performed equivalently or, at times, better than the other versions of the VisAWI. This was also evident in the results from the EFAs, where most of the VisAWI-pos items loaded as expected given a four-factor solution (ie, items loading onto distinct factors depending on their intended facet). In contrast, multiple items for the original VisAWI showed problematic loadings, with most of those problematic items (4/6) being negatively formulated.

The fact that the reverse-coded items caused problems regarding the VisAWI's psychometric quality, such as in the EFAs, is not surprising, given past research findings. As summarized in Subsection 2.3, reverse-coded items not only negatively affect the psychometric quality of a survey scale but can also lead to misresponses because their wording leaves room for interpretation, thus causing comprehension issues among participants. Given that the VisAWI-pos avoids reverse-coded items, researchers need not worry about these issues related to the comprehension of item wording.

In summary, the VisAWI-pos proved to be a superior alternative to the English VisAWI, delivering comparable results to the original while coming with improved psychometric quality and avoiding the drawbacks of reverse-coded items.

6.2. Psychometric evaluation of the VisAWI and VisAWI-S

As mentioned above, the original VisAWI and VisAWI-S papers (Moshagen & Thielsch, 2013, 2010) worked with a German sample and a German version of the scale. Therefore, no psychometric evaluation of the English VisAWI versions with an English-speaking sample has taken place to date (Abbas et al., 2022). The present work thus presents the first psychometric evaluation for the English versions of the VisAWI. Concerning both the VisAWI and the VisAWI-S, the results show evidence for good psychometric properties of the scales, both in terms of reliability (ie, internal consistency coefficients α and ω) and validity (ie, CFA, distinguishing between websites, correlations with convergent and divergent scales). Fit indices for the VisAWI-S mainly favored the model, with only the $RMSEA$ above the desired cutoffs (both > 0.06 and > 0.08). For the VisAWI, fit statistics were close to the ideal cutoffs recommended by Hu and Bentler (1999) but mostly outside the desired thresholds. In addition, likewise to the VisAWI-pos, the English version of the VisAWI and the VisAWI-S are of comparable psychometric quality to other previously published versions of the VisAWI in German, Farsi, and Arabic.

Based on the results presented here, the English versions of the VisAWI and VisAWI-S are of acceptable psychometric quality and comparable to other versions. Nevertheless, as highlighted above, the VisAWI-pos showed even more favorable psychometric results and should thus be considered the preferable alternative. Furthermore, results from the EFA showed that the negative items within the original VisAWI cause problems, deviating from the expected theoretical structure of the scale. To avoid these issues, researchers should work with the VisAWI-pos instead of the original VisAWI or use the VisAWI-S if more suitable to their needs.

6.3. Room for improvement with craftsmanship

While all three VisAWI versions performed relatively well in the analyses, there were multiple issues with the craftsmanship facet. A negative variance (ie, Heywood case) for craftsmanship was encountered in the VisAWI-pos' CFA, which was also documented for the AR-VisAWI (Abbas et al., 2022). Furthermore, in the EFAs, the craftsmanship items showed the least favorable loadings, both for the VisAWI and the VisAWI-pos. The researchers see three probable reasons for this misfit of the craftsmanship facet.

First, there has been a change in website design over the last decade. Goree et al. (2021) showed that website design has become more similar since 2007, especially regarding page layout. This homogenization of design might be a factor that influences how individuals judge a website and what aspects have to be considered in website design. Thus, what makes a website "well crafted" has likely changed over the last decade since the VisAWI's original development. Interestingly, the other three facets of the VisAWI-pos have held up well despite these changes in website design.

Second, multiple indicators suggest that craftsmanship is an overarching construct, closely intertwined and not neatly distinguishable from the other facets of the VisAWI. When fitting a CFA model without a higher-order factor for the VisAWI items, very high correlations between craftsmanship and simplicity (0.93) and craftsmanship and diversity (0.91) were observed. For the VisAWI-pos, these correlations of craftsmanship with simplicity (0.93) and diversity (0.89) were also very high. Abbas et al. (2022) reported on the same issue with the AR-VisAWI. Furthermore, items for craftsmanship frequently cross-loaded onto the same factors as items for simplicity and diversity in the EFAs, showing that craftsmanship was not neatly distinguishable from these other factors. The VisAWI's theoretical model, especially craftsmanship and its relation to the other facets, thus appears to differ from what Moshagen and Thielsch (2010) suggested.

Third, participants' cultural backgrounds possibly influenced the craftsmanship ratings. No Heywood cases were reported for the original German VisAWI validation studies, but they were present in Arabic and English samples. Thus, the VisAWI's original model might apply differently to non-German samples. Hence, further research is needed concerning the cultural measurement invariance of the VisAWI and its underlying theoretical model in versions other than German.

Regardless, while craftsmanship was an issue for both versions of the VisAWI investigated here, original and positive, the results favor the VisAWI-pos regarding the craftsmanship facet. In the present work, the craftsmanship items of the original VisAWI showed no loadings > 0.20 to a unique factor in the four-factor EFA. In contrast, three of the four craftsmanship items for the VisAWI-pos showed loadings > 0.30 on a fourth factor, albeit with high cross-loadings. Thus, the craftsmanship facet was more pronounced in the VisAWI-pos but should be improved or reconsidered for both scales. Regarding the Heywood case, while the VisAWI exhibited no negative residual variance for craftsmanship in the second study, such a Heywood case was also reported by Abbas et al. (2022) while developing the AR-VisAWI. Hence, the negative residual variances and the issues with craftsmanship are also present in the model of other VisAWI versions and are not endemic to the VisAWI-pos. Furthermore, the Heywood case should not be seen as a substantial disadvantage of the VisAWI-pos over the VisAWI, given that the negative residual variance for the craftsmanship factor of the VisAWI-pos was minimal (−0.003) and of comparable magnitude to the positive residual variance of craftsmanship for the original VisAWI (0.002). Even when fixing the problematic loading for craftsmanship to prevent the Heywood case from occurring–a suggested remedy for such cases (Chen et al., 2001; Farooq, 2022)–the model fit and loadings did not change substantially.

In summary, the facet of craftsmanship and the VisAWI's theoretical model requires further investigation. Nevertheless, given the otherwise favorable results, the VisAWI-pos can still be used, especially to measure the other three facets of aesthetics, until such an investigation has been completed.

6.4. Using the VisAWI-pos

Given that the VisAWI-pos is based on the original VisAWI, researchers should follow the guidelines on using the VisAWI (Thielsch & Moshagen, 2015) when working with the VisAWI-pos. Items should be presented alongside the original VisAWI instructions, and responses are collected using a 7-point Likert-type scale from 1 (strongly disagree) to 7 (strongly agree). Figure A4 in the Appendix contains the complete VisAWI-pos, including instructions and answering options. After data collection, items are scored like the original VisAWI by calculating mean values across the respective items of the four facets. The overall score is calculated using the mean value across the facet scores. No item reversal is needed before calculating mean scores. The VisAWI-pos was developed in a desktop computer setting, like the original VisAWI. Thus, researchers can use it confidently to judge a desktop website's visual aesthetics. It is recommended that when using the VisAWI-pos in other settings (eg, mobile devices), the psychometric quality of the VisAWI-pos is investigated before interpreting the collected data. Finally, when reporting their results, researchers should cite the original work on the VisAWI (Thielsch & Moshagen, 2015) in addition to citing the present paper, given that the VisAWI-pos is closely based on the original VisAWI.

6.5. Limitations

First, crowd-sourced samples were used for the two user studies. Thus, future research should investigate if the VisAWI-pos also works with samples from other populations, such as students or volunteer samples. Despite this, past work has shown that crowd-sourced samples are comparable in quality to other samples (Buhrmester et al., 2011) and that the crowdsourcing platform used (ie, Prolific) is preferable to alternative providers (Peer et al., 2022). Second, participants in both studies were based either in the USA or the UK. Further investigation is thus needed to see if the VisAWI-pos works with participants from other English-speaking countries and if the scale can be used with non-native speakers. Third, while the set of websites used as stimuli in study two was carefully selected to represent a broad spectrum of content areas, the selection process purposefully aimed for sites with which participants were unfamiliar. It thus remains to be seen how the VisAWI-pos performs with other types of websites, especially those with which participants are familiar. Similarly, the tasks used only had participants interact with the site over a limited time frame and not, for example, throughout multiple sessions. In genuine usage, participants will often return to specific sites they use. It thus remains to be seen if the VisAWI-pos can measure aesthetics over time. Finally, past research has shown that users' experiences differ depending on their usage mode (Iten et al., 2018). While a broad range of tasks was created, these were not purposefully varied to create groups with different usage modes. Furthermore, given the crowd-sourced sample, most participants were likely in a work-related mindset and thus in "goal mode" (Hassenzahl et al., 2002; Iten et al., 2018). Therefore, further clarification is needed on how the VisAWI-pos performs in tasks where usage mode varies and where users are in a more leisure-oriented mindset.

6.6. Future work

As mentioned above, future work should look at the craftsmanship facet, investigating how it relates to the other facets of the VisAWI model and if it truly is a distinct facet that can be measured using the proposed items of either the VisAWI or VisAWI-pos. Second, while results from the present evaluation supported the psychometric quality of the VisAWI-pos, future work should look at the measurement invariance of the scale, seeing if the scale's quality remains robust over different samples and subgroups. Third, the present work only focused on the English version of the VisAWI. Thus, future work should create alternative positive items for the other language versions of the VisAWI so that researchers can avoid issues with negative item wording when using those versions of the scale. Finally, both studies were situated in a desktop website setting. Hence, future work should consider whether the VisAWI-pos and the English versions of the VisAWI and VisAWI-S also hold up with mobile devices or other non-desktop website interfaces.

7. Conclusion

Throughout two studies, an alternative version of the VisAWI, which does not use reverse-coded items, was created and validated. The resulting scale, called VisAWI-pos, provided almost identical results to the original VisAWI while exhibiting equal or better psychometric quality. Furthermore, as part of these endeavors, the English version of the VisAWI and VisAWI-S were also validated for the first time, showing that both versions are comparable to other language versions of the scale, which have already seen proper psychometric evaluation. Finally, the present work identified issues related to the VisAWI's craftsmanship facet, which require further clarification. Overall, the present work provides researchers with a new survey scale for measuring aesthetics in the context of websites, which builds upon the established VisAWI while avoiding issues that come with the use of negatively formulated items in the original scale.

Notes

1. Indicated by the high citation count on Google Scholar for the papers on the original VisAWI (662) and the VisAWI-S (143), retrieved on 30. January 2023.
2. The original task for stimuli website four was incorrectly answered by all three participants, while two out of three wrongly responded to the first version of the task for website nine.
3. In the pre-registration, the cutoff for the CFI was set at 0.98 based on a citation of Hu and Bentler (1999) in Steinmetz (2015). Checking again with the original source, the actual value recommended by Hu and Bentler (i.e., 0.95) was settled on because it was always the intention to follow their recommendations.
4. Fit statistics reported in Moshagen and Thielsch (2010): $\chi^2(131) = 405.57$, $p < 0.01$, $RMSEA = 0.064$, $SRMR = 0.052$, $CFI = 0.945$.
5. Fit statistics reported in Abbas et al. (2022): $\chi^2(86) = 196.60$, $RMSEA = 0.08$, $SRMR = 0.05$, $CFI = 0.94$.
6. Fit statistics reported in Saremi et al. (2023): $RMSEA = 0.048$, $SRMR = 0.055$, $CFI = 0.916$.

Acknowledgments

Special thanks to Anthony Mahler for their language expertise concerning the initial development of the VisAWI-pos items.

Author contributions

SP and MH developed the initial item set with inputs from FB and KO. SP, MH, and FB contributed to the conception and design of the item development and first study. SP and MH performed the statistical analysis for the first study. SP, NF, and FB designed the second study with inputs from KO. SP and NF developed the set of stimuli websites and tasks for study two with inputs from FB. SP and NF performed the statistical analysis of the second study. SP wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Ethical approval

Ethical approval was not needed, given the non-interventional nature of both studies, based on local and institutional guidelines. However, the large sample study two was reviewed and approved by the ethics committee of the Faculty of Psychology of the University Basel (approval number 007-22-01), while no review was undertaken for the smaller sample study one. In both studies, all respondents gave informed consent before proceeding with the online survey.

Disclosure statement

This work is financed entirely by the authors' research group, as they received no additional funding. The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID

Sebastian A. C. Perrig http://orcid.org/0000-0002-4301-8206
Nick von Felten http://orcid.org/0000-0003-0278-9896
Marimo Honda http://orcid.org/0009-0000-6599-6556
Klaus Opwis http://orcid.org/0000-0003-0509-8070
Florian Brühlmann http://orcid.org/0000-0001-8945-3273

Data availability statement

The data, complete analysis (R scripts), and supplementary materials (printouts of online surveys) for this study are available on OSF (https://osf.io/efy4s).

References

Abbas, A., Hirschfeld, G., & Thielsch, M. T. (2022). An Arabic version of the visual aesthetics of websites inventory (AR-VisAWI): Translation and psychometric properties. *International Journal of Human–Computer Interaction*, *39*(14), 2785–2795. https://doi.org/10.1080/10447318.2022.2085409

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 2689–2698). Association for Computing Machinery. https://doi.org/10.1145/1978942.1979336

Barnette, J. J. (2000). Effects of stem and likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361–370. https://doi.org/10.1177/00131640021970592

Bhandari, U., Chang, K., & Neben, T. (2019). Understanding the impact of perceived visual aesthetics on user evaluations: An emotional perspective. *Information & Management*, *56*(1), 85–93. https://doi.org/10.1016/j.im.2018.07.003

Brooke, J. (1996). SUS: A'quick and dirty' usability scale. In B. Thomas, B. Weerdmeester, I. L. McClelland, & P. W. Jordan (Eds.), *Usability evaluation in industry*, (Vol. 189, pp. 189–194). CRC Press.

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022. https://doi.org/10.1016/j.metip.2020.100022

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468–508. https://doi.org/10.1177/0049124101029004003

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Dalal, D. K., & Carter, N. T. (2014). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (1st ed., pp. 112–132). Routledge.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed., Vol. 26). SAGE Publications, Inc.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Dodeen, H. (2023). The effects of changing negatively worded items to positively worded items on the reliability and the factor structure of psychological scales. *Journal of Psychoeducational Assessment*, *41*(3), 298–310. https://doi.org/10.1177/07342829221141934

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden [statistics and research methods]* (5th ed.). Beltz.

Farooq, R. (2022). Heywood cases: Possible causes and solutions. *International Journal of Data Analysis Techniques and Strategies*, *14*(1), 79–88. https://doi.org/10.1504/IJDATS.2022.121506

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology* (1st ed.). SAGE Publications, Ltd.

George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: A simple guide and reference* (16th ed.). Routledge.

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg self-esteem scale. *Assessment*, *27*(2), 404–418. https://doi.org/10.1177/1073191117746503

Goree, S., Doosti, B., Crandall, D., & Su, N. M. (2021). Investigating the homogenization of web design: A mixed-methods approach. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, *19*(4), 319–349. https://doi.org/10.1207/s15327051hci1904_2

Hassenzahl, M., Kekez, R., Burmester, M. (2002). The importance of a software's pragmatic quality depends on usage modes. In *Proceedings of the 6th International Conference on Work with Display Units (WWDU 2002)* (pp. 275–276).

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction*, *25*(3), 235–260. https://doi.org/10.1080/07370024.2010.500139

Hausman, A. V., & Siekpe, J. S. (2009). The effect of web interface features on consumer online purchase intentions. *Journal of Business Research*, *62*(1), 5–13. https://doi.org/10.1016/j.jbusres.2008.01.018

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics – Theory and Methods*, *19*(10), 3595–3617. https://doi.org/10.1080/03610929008830400

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society of London*, *134*(824), 486–501. https://doi.org/10.1098/rspa.1931.0209

Hirschfeld, G., & Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. *Ergonomics*, *58*(2), 310–320. https://doi.org/10.1080/00140139.2014.965228

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Iten, G. H., Troendle, A., & Opwis, K. (2018). Aesthetics in context—the role of aesthetics and usage mode for a website's success. *Interacting with Computers*, *30*(2), 133–149. https://doi.org/10.1093/iwc/iwy002

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kam, C. C. S., Meyer, J. P., & Sun, S. (2021). Why do people agree with both regular and reversed items? A logical response perspective. *Assessment*, *28*(4), 1110–1124. https://doi.org/10.1177/10731911211001931

Kam, C. C. S., & Sun, S. (2022). Method factor due to the use of reverse-keyed items: Is it simply a response style artifact? *Current Psychology*, *41*(3), 1204–1212. https://doi.org/10.1007/s12144-020-00645-z

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kortum, P., Acemyan, C. Z., & Oswald, F. L. (2021). Is it time to go positive? Assessing the positively worded system usability scale (SUS). *Human Factors*, *63*(6), 987–998. https://doi.org/10.1177/0018720819881556

Kurosu, M., & Kashimura, K. (1995). *Apparent usability vs. inherent usability: Experimental analysis on the determinants of the apparent usability* [Paper presentation]. Conference Companion on Human Factors in Computing Systems, New York, NY, USA (pp. 292–293). Association for Computing Machinery. https://doi.org/10.1145/223355.223680

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and usability for education and work* (pp. 63–76). Springer.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, *60*(3), 269–298. https://doi.org/10.1016/j.ijhcs.2003.09.002

Lee, S., & Koubek, R. J. (2010). Understanding user preferences based on usability and aesthetics before and after actual use. *Interacting with Computers*, *22*(6), 530–543. https://doi.org/10.1016/j.intcom.2010.05.002

Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the system usability scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). *UMUX-LITE: When there's no time for the SUS* [Paper presentation]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA (p. 2099–2102). Association for Computing Machinery.

Lindgaard, G. (2007). Aesthetics, visual appeal, usability and user satisfaction: What do the user's eyes tell the user's brain? *Australian Journal of Emerging Technologies & Society*, *5*(1), 1–14.

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., & Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction*, *18*(1), 1–30. https://doi.org/10.1145/1959022.1959023

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*(2), 196–204. https://doi.org/10.1080/00223891.2011.645936

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530. https://doi.org/10.1093/biomet/57.3.519

McDonald, R. P. (1999). *Test theory: A unified treatment* (1st ed.). Psychology Press.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Minge, M., & Thüring, M. (2018). Hedonic and pragmatic halo effects at early stages of user experience. *International Journal of Human-Computer Studies*, *109*(3), 13–25. https://doi.org/10.1016/j.ijhcs.2017.07.007

Moshagen, M., & Thielsch, M. (2013). A short version of the visual aesthetics of websites inventory. *Behaviour & Information Technology*, *32*(12), 1305–1311. https://doi.org/10.1080/0144929X.2012.694910

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). Mcgraw hill book company.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pengnate, S. F., Sarathy, R., & Lee, J. (2019). The engagement of website initial aesthetic impressions: An experimental investigation. *International Journal of Human–Computer Interaction*, *35*(16), 1517–1531. https://doi.org/10.1080/10447318.2018.1554319

Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). *Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI* [Paper presentation]. Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3544549.3585808

Perrig, S. A. C., Ueffing, D., Opwis, K., & Brühlmann, F. (2023). Smartphone app aesthetics influence users' experience and performance. *Frontiers in Psychology*, *14*, 1113842. https://doi.org/10.3389/fpsyg.2023.1113842

Pettersson, I., Lachner, F., Frison, A.-K., Riener, A., & Butz, A. (2018). *A bermuda triangle? a review of method application and triangulation in user experience evaluation* [Paper presentation]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, New York, NY, USA (p. 1–16). Association for Computing Machinery.

Pilotte, W. J., & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, *50*(3), 603–610. https://doi.org/10.1177/0013164490503016

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), 46–54, 124.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013). *Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness* [Paper presentation]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA (pp. 2049–2058). Association for Computing Machinery.

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192–199. https://doi.org/10.7334/psicothema2014.266

Saremi, M., Sadeghi, V., Khodakarim, S., & Maleki-Ghahfarokhi, A. (2023). Farsi version of visual aesthetics of website inventory (FV-VisAWI): Translation and psychometric evaluation. *International Journal of Human–Computer Interaction*, *39*(4), 834–841. https://doi.org/10.1080/10447318.2022.2049138

Sauro, J., & Lewis, J. R. (2011). *When designing usability questionnaires, does it hurt to be positive?* [Paper presentation]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA (pp. 2215–2224). Association for Computing Machinery. https://doi.org/10.1145/1978942.1979266

Scharowski, N., & Perrig, S. A. C. (2023). Distrust in (X)AI – measurement artifact or distinct construct? *CHI 2023 TRAIT Workshop on Trust and Reliance in AI-Human Teams*. CHI.

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schrepp, M., Otten, R., Blum, K., & Thomaschewski, J. (2021). What causes the dependency between perceived aesthetics and perceived usability? *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(6), 78–85. https://doi.org/10.9781/ijimai.2020.12.005

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement*, *41*(4), 1101–1114. https://doi.org/10.1177/001316448104100420

Seckler, M., Opwis, K., & Tuch, A. N. (2015). Linking objective design factors with subjective aesthetics: An experimental study on how structure and color of websites affect the facets of users' visual aesthetic perception. *Computers in Human Behavior*, *49*(C), 375–389. https://doi.org/10.1016/j.chb.2015.02.056

Seng, T. L., & Mahmoud, M. A. S. (2020). Perceived e-service quality and e-store loyalty: The moderated mediating effect of webpage aesthetics and e-customer satisfaction. *International Journal of Advanced and Applied Sciences*, *7*(5), 111–117. https://doi.org/10.21833/ijaas.2020.05.014

Seo, K.-K., Lee, S., Chung, B. D., & Park, C. (2015). Users' emotional valence, arousal, and engagement based on perceived usability and aesthetics for web sites. *International Journal of Human-Computer Interaction*, *31*(1), 72–87. https://doi.org/10.1080/10447318.2014.959103

Skulmowski, A., Augustin, Y., Pradel, S., Nebel, S., Schneider, S., & Rey, G. D. (2016). The negative impact of saturation on website trustworthiness and appeal: A temporal model of aesthetic website perception. *Computers in Human Behavior*, *61*, 386–393. https://doi.org/10.1016/j.chb.2016.03.054

Steinmetz, H. (2015). *Lineare Strukturgleichungsmodelle. Eine Einführung mit R [linear structural equation modeling. an introduction with R]* (2nd ed.). Rainer Hampp Verlag.

Stewart, T. J., & Frye, A. W. (2004). Investigating the use of negatively phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine*, *79*(10 Suppl), S18–S20. https://doi.org/10.1097/00001888-200410001-00006

Suárez Álvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Thielsch, M. T., & Hirschfeld, G. (2021). Quick assessment of web content perceptions. *International Journal of Human–Computer Interaction*, *37*(1), 68–80. https://doi.org/10.1080/10447318.2020.1805877

Thielsch, M. T., Moshagen, M. (2015). *VisAWI manual*. Retrieved February 27, 2023, from https://visawi.uid.com/pdf/VisAWI_Manual_EN.pdf

Thielsch, M. T., Scharfen, J., Masoudi, E., & Reuter, M. (2019). Visual aesthetics and performance: A first meta-analysis. *Proceedings of Mensch Und Computer 2019* (pp. 199–210). Association for Computing Machinery.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145. https://doi.org/10.1016/S0953-5438(00)00031-X

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior, 28*(5), 1596–1607. https://doi.org/10.1016/j.chb.2012.03.024

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLOS One*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Venta, A., Bailey, C. A., Walker, J., Mercado, A., Colunga-Rodriguez, C., Ángel González, M., & Dávalos-Picazo, G. (2022). Reverse-coded items do not work in Spanish: Data from four samples using established measures. *Frontiers in Psychology*, *13*, 828037. https://doi.org/10.3389/fpsyg.2022.828037

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology*, *8*, 84–136. https://doi.org/10.2307/270754

Whittaker, T. A., & Schumacker, R. E. (2022). *A beginner's guide to structural equation modeling* (5th ed.). Routledge.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). 06) Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, *30*(1), 72–91. https://doi.org/10.1086/374697

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PLOS One*, *11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

About the authors

Sebastian A. C. Perrig is a Ph.D. candidate and research assistant in the Human-Computer Interaction research group at the Center for Cognitive Psychology and Methodology at the University of Basel, Switzerland. His research interests include user and player experience, scale development and validation, visual aesthetics, and online survey research.

Nick von Felten is an MSc student and student assistant in the Human-Computer Interaction research group at the Center for Cognitive Psychology and Methodology at the University of Basel, Switzerland. His research interests include explainable artificial intelligence, questionnaire development, statistical methods, and user experience.

Marimo Honda received her BSc degree in psychology from the University of Basel, Switzerland, in 2022. Currently, she is an MSc student in psychology specializing in Human-Computer Interaction at the same university, with a particular interest in user experience and visual aesthetics.

Klaus Opwis was appointed head of the Center for General Psychology and Methodology at the University of Basel, Switzerland, in 1995. His research interests include applied cognitive psychology, visual aesthetics, research methods, and HCI research.

Florian Brühlmann is the research director of the Human-Computer Interaction research group at the Center for General Psychology and Methodology at the University of Basel, Switzerland. His research interests include user and player experience, human-AI interaction, and research methods in HCI.

Appendix

Table A1. Pearson's correlations from study two of the VisAWI (top) and VisAWI-pos (bottom) facets with the convergent and divergent scales.

VisAWI	Simplicity	Diversity	Colorfulness	Craftsmanship
Diversity	0.69			
Colorfulness	0.71	0.73		
Craftsmanship	0.80	0.81	0.75	
VisAWI Total Score	0.89	0.90	0.88	0.93
Classic aesthetics	0.82	0.71	0.71	0.76
Expressive aesthetics	0.56	0.80	0.59	0.65
Attractiveness (UEQ)	0.77	0.76	0.72	0.77
Hedonic quality (UEQ)	0.78	0.55	0.57	0.65
Pragmatic quality (UEQ)	0.57	0.82	0.62	0.68
Usability (SUS score)	0.76	0.53	0.55	0.61
Web-CLIC-S	0.77	0.64	0.65	0.69
NPS (raw values)	0.63	0.66	0.59	0.64
VisAWI-pos	**Simplicity (pos)**	**Diversity (pos)**	**Colorfulness (pos)**	**Craftsmanship (pos)**
Diversity (pos)	0.67			
Colorfulness (pos)	0.68	0.75		
Craftsmanship (pos)	0.83	0.80	0.75	
VisAWI-pos Total Score	0.88	0.90	0.88	0.94
Classic aesthetics	0.84	0.72	0.72	0.80
Expressive aesthetics	0.55	0.82	0.65	0.66
Attractiveness (UEQ)	0.77	0.77	0.74	0.78
Hedonic quality (UEQ)	0.78	0.56	0.57	0.68
Pragmatic quality (UEQ)	0.53	0.82	0.67	0.66
Usability (SUS score)	0.78	0.54	0.56	0.66
Web-CLIC-S	0.78	0.66	0.66	0.73
NPS (raw values)	0.62	0.67	0.60	0.65

All correlations were significant at $p < 0.0001$.

Table A2. Measurement-error-free correlations from study two of the VisAWI versions with the convergent and divergent scales, calculated using CFA.

	VisAWI	VisAWI-pos	VisAWI-S
Classic aesthetics	0.91	0.93	0.92
Expressive aesthetics	0.83	0.83	0.81
Attractiveness (UEQ)	0.91	0.91	0.91
Hedonic quality (UEQ)	0.86	0.85	0.84
Pragmatic quality (UEQ)	0.78	0.79	0.77
Usability (SUS score)	0.70	0.73	0.69
Web-CLIC-S	0.87	0.89	0.87
NPS (raw values)	0.73	0.74	0.72

All correlations were significant at $p < 0.0001$.

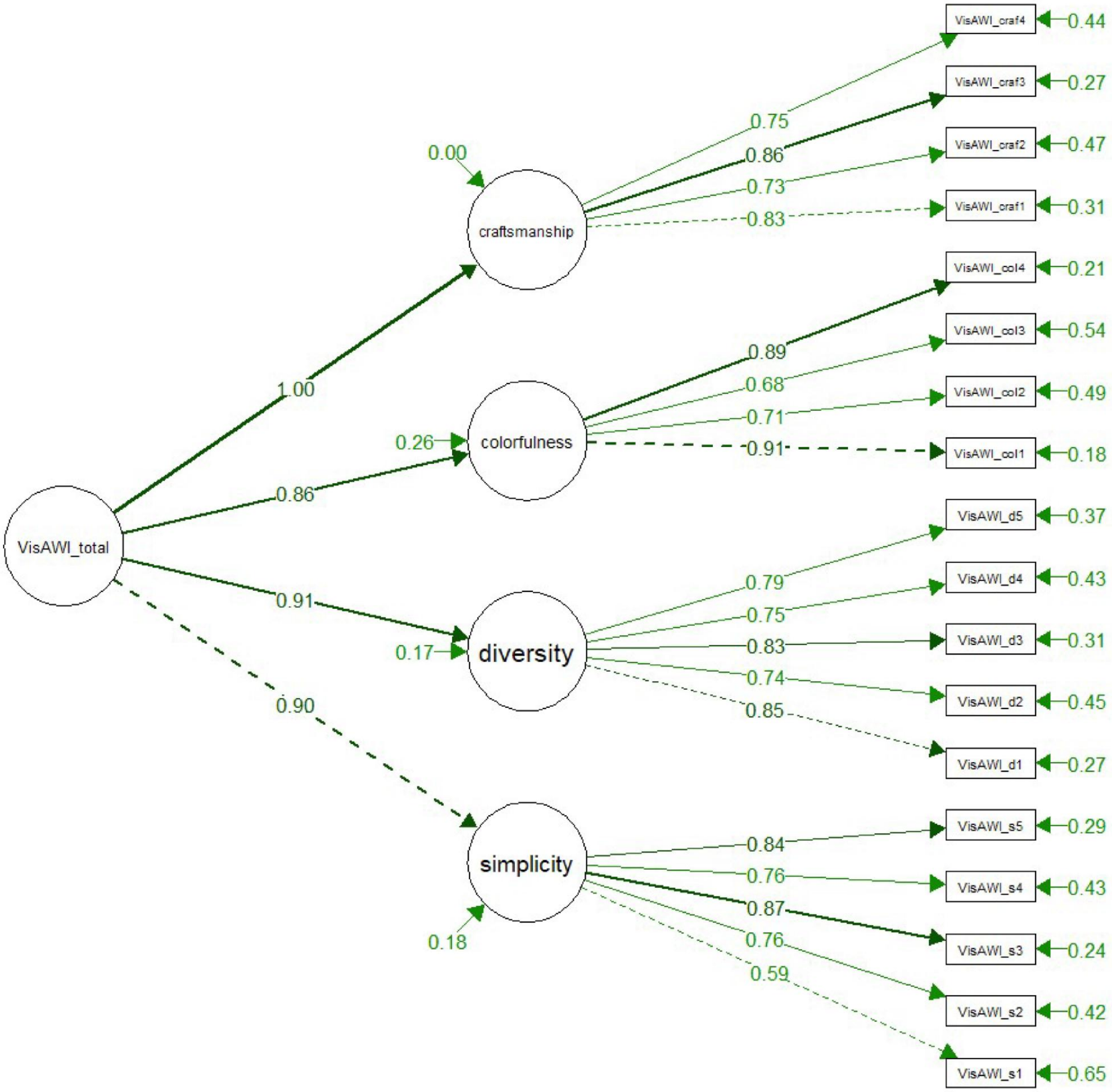


Figure A1. Measurement model used in the CFA for the original VisAWI in study two, including loadings. Dotted lines indicate loadings that were constrained to one.

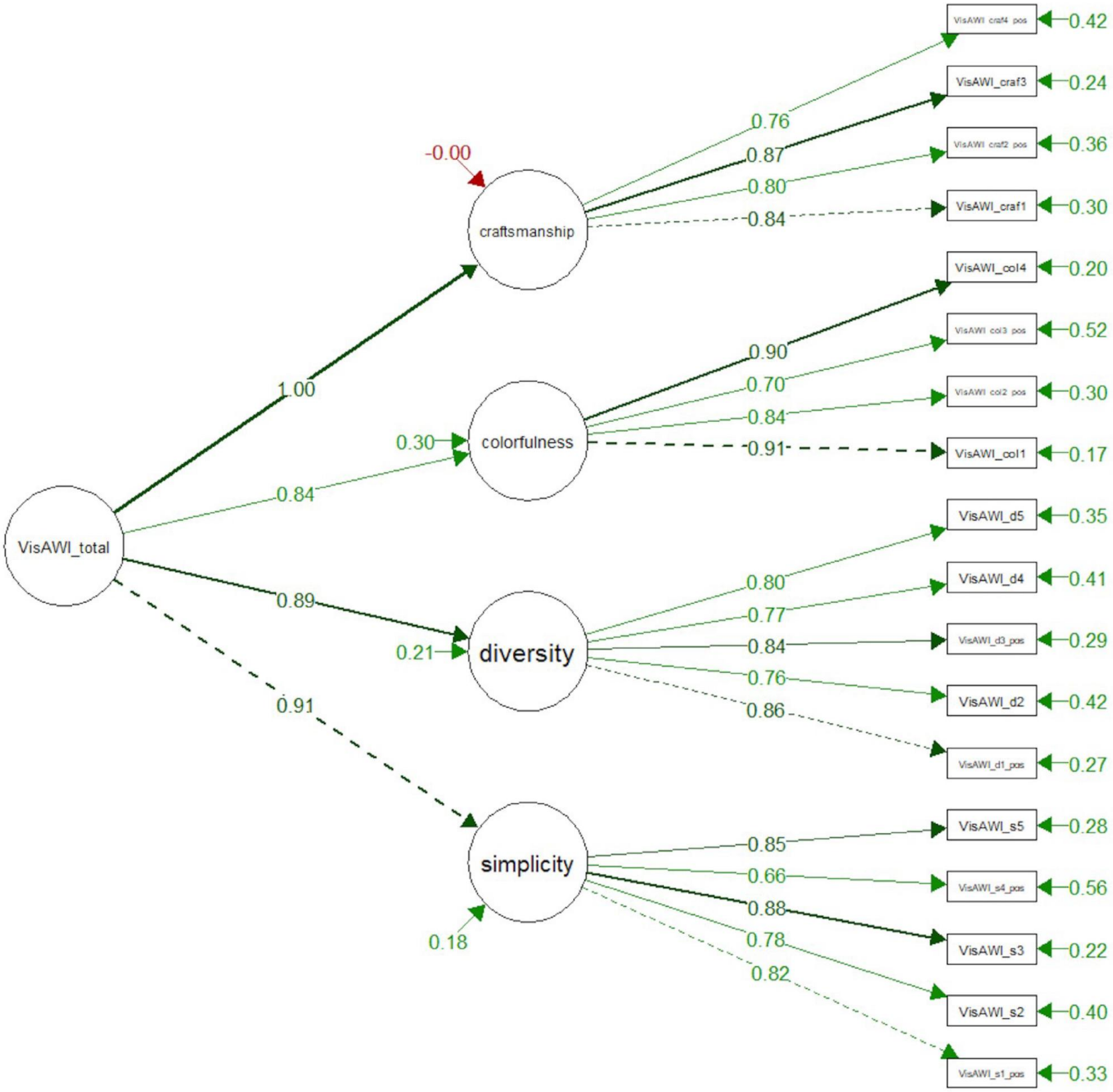


Figure A2. Measurement model used in the CFA for the VisAWI-pos in study two, including loadings. Dotted lines indicate loadings that were constrained to one.

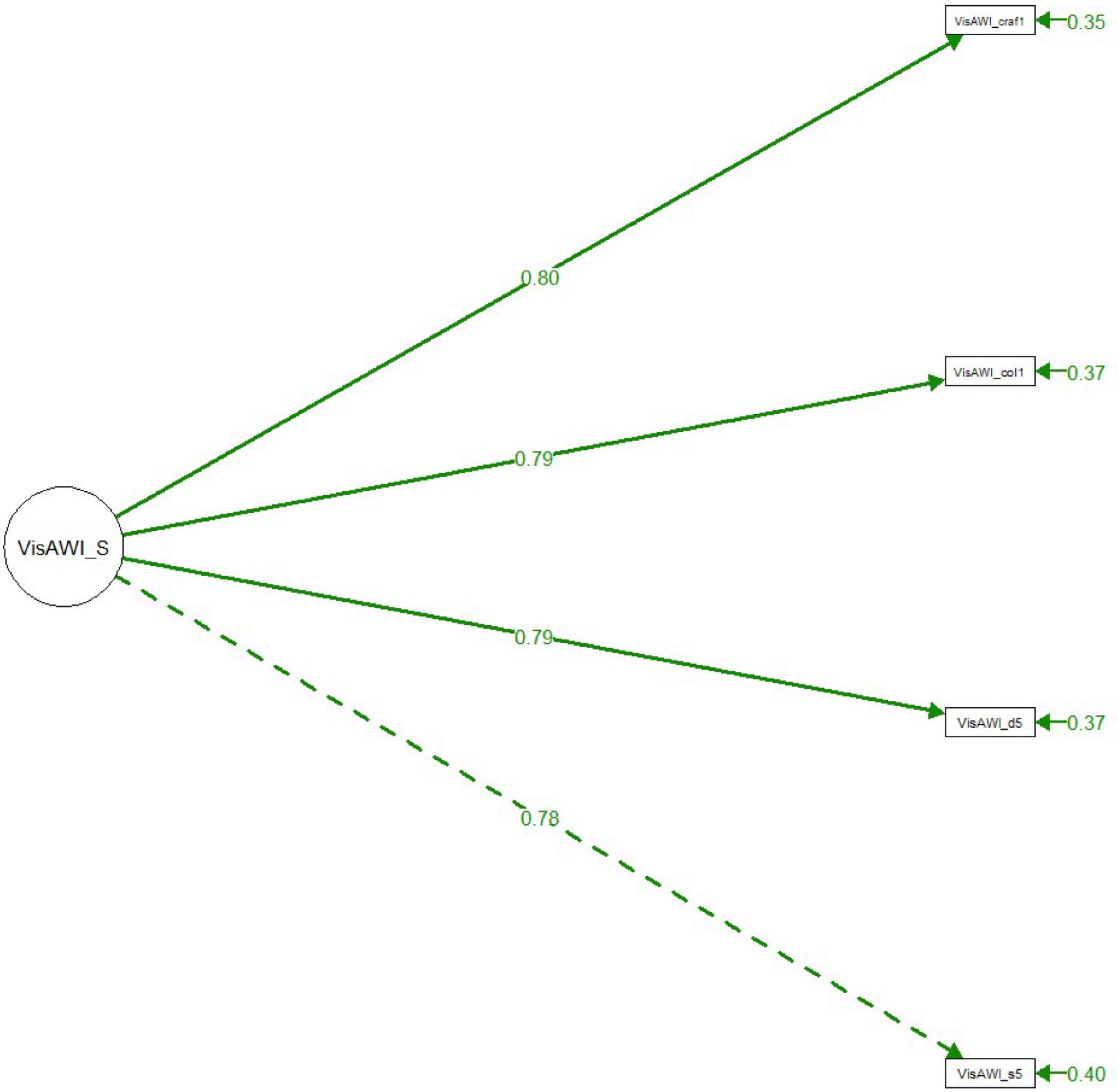
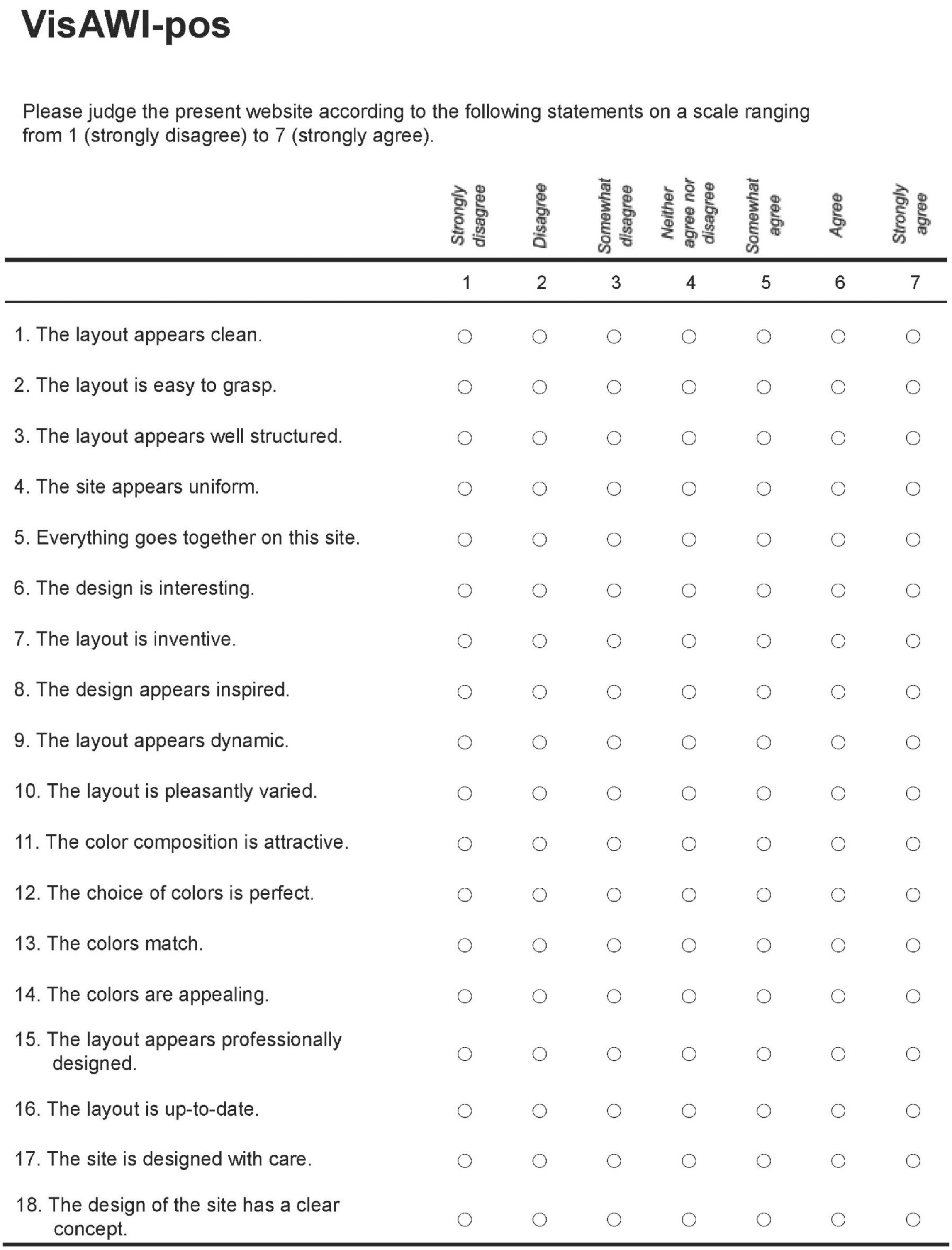


Figure A3. Measurement model used in the CFA for the VisAWI-S in study two, including loadings. Dotted lines indicate loadings that were constrained to one.



VisAWI-pos

Please judge the present website according to the following statements on a scale ranging from 1 (strongly disagree) to 7 (strongly agree).

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
	1	2	3	4	5	6	7
1. The layout appears clean.	○	○	○	○	○	○	○
2. The layout is easy to grasp.	○	○	○	○	○	○	○
3. The layout appears well structured.	○	○	○	○	○	○	○
4. The site appears uniform.	○	○	○	○	○	○	○
5. Everything goes together on this site.	○	○	○	○	○	○	○
6. The design is interesting.	○	○	○	○	○	○	○
7. The layout is inventive.	○	○	○	○	○	○	○
8. The design appears inspired.	○	○	○	○	○	○	○
9. The layout appears dynamic.	○	○	○	○	○	○	○
10. The layout is pleasantly varied.	○	○	○	○	○	○	○
11. The color composition is attractive.	○	○	○	○	○	○	○
12. The choice of colors is perfect.	○	○	○	○	○	○	○
13. The colors match.	○	○	○	○	○	○	○
14. The colors are appealing.	○	○	○	○	○	○	○
15. The layout appears professionally designed.	○	○	○	○	○	○	○
16. The layout is up-to-date.	○	○	○	○	○	○	○
17. The site is designed with care.	○	○	○	○	○	○	○
18. The design of the site has a clear concept.	○	○	○	○	○	○	○

Figure A4. The complete VisAWI-pos, including instructions and response options, for use in a study.

Transparency in Measurement Reporting: A Systematic Literature Review of CHI PLAY

LENA FANYA AESCHBACH, SEBASTIAN A. C. PERRIG, LORENA WEDER, KLAUS OPWIS, and FLORIAN BRÜHLMANN, Department for General Psychology and Methodology, University of Basel, Switzerland

Measuring theoretical concepts, so-called constructs, is a central challenge of Player Experience research. Building on recent work in HCI and psychology, we conducted a systematic literature review to study the transparency of measurement reporting. We accessed the ACM Digital Library to analyze all 48 full papers published at CHI PLAY 2020, of those, 24 papers used self-report measurements and were included in the full review. We assessed specifically, whether researchers reported *What*, *How* and *Why* they measured. We found that researchers matched their measures to the construct under study and that administrative details, such as number of points on a Likert-type scale, were frequently reported. However, definitions of the constructs to be measured and justifications for selecting a particular scale were sparse. Lack of transparency in these areas threaten the validity of singular studies, but further compromise the building of theories and accumulation of research knowledge in meta-analytic work. This work is limited to only assessing the current transparency of measurement reporting at CHI PLAY 2020, however we argue this constitutes a fair foundation to assess potential pitfalls. To address these pitfalls, we propose a prescriptive model of a measurement selection process, which aids researchers to systematically define their constructs, specify operationalizations, and justify why these measures were chosen. Future research employing this model should contribute to more transparency in measurement reporting. The research was funded through internal resources. All materials are available on https://osf.io/4xz2v/.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **HCI theory, concepts and models**.

Additional Key Words and Phrases: Transparency, Measurement, Literature review, Methodology, Validity, Survey, Open Science

ACM Reference Format:
Lena Fanya Aeschbach, Sebastian A. C. Perrig, Lorena Weder, Klaus Opwis, and Florian Brühlmann. 2021. Transparency in Measurement Reporting: A Systematic Literature Review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 233 (October 2021), 21 pages. https://doi.org/10.1145/3474660

1 INTRODUCTION

One of the most common ways to collect data in player experience (PX) research is to employ the use of self-report, survey-based instruments. These measurements function by presenting participants with short, declarative statements (e.g., "I enjoyed playing this game very much") or questions followed with a response scale (for example, with options from *not at all* to *very much*, *never* to *often*, or with a numerical range). The type of data collected through this process is both subjective and

Authors' address: Lena Fanya Aeschbach, lena.aeschbach@unibas.ch; Sebastian A. C. Perrig, sebastian.perrig@unibas.ch; Lorena Weder, lorena.weder@unibas.ch; Klaus Opwis, klaus.opwis@unibas.ch; Florian Brühlmann, Department for General Psychology and Methodology, University of Basel, Missionsstrasse 62a, Basel, Switzerland, 4055, florian.bruehlmann@unibas.ch.

2573-0142/2021/9-ART233. https://doi.org/10.1145/3474660

quantitative. Subjective meaning, participants are asked to give an honest answer from their own perspective to a question or statement (usually referred to as an *item*). Quantitative, on the other hand, refers to the form of the data the researcher will receive which is numerical. Because of this, scales and questionnaires offer a multitude of benefits to researchers [10, 13]. The numerical data can be used for a variety of statistical analyses, while the subjectivity allows to measure complex psychological processes which are not directly observable. For example, motivation or identification, which are both constructs that must be inferred from observed variables. This 'tapping into the subjective' is often necessary to understand the intricate interactions between the players and the games studied. Further, using a valid questionnaire and experimental design, the numerical data collected can be used for statistical analyses that can provide evidence of a cause and effect relationship. When using a reliable questionnaire, the differences between experimental groups on this questionnaire may be attributed to the experiment, rather than errors of measurement. In addition, numerical data can be shaped and displayed in a variety of visual representations. This also extends to allow for easier interpretable estimates of magnitude and comparisons of importance.

What is being measured using these measurements are referred to as constructs. A construct has been defined as a theoretical instance of a concept, meaning it is an attribute or idea that is not directly observable or measurable [26]. However, the various benefits associated with measuring constructs using survey scales can only be achieved when we are certain that the questionnaire or scale we employ measures what we have set out to measure. Therefore, when attempting to make a causal statement about an unobservable construct after an experiment, for example motivation, one must be certain that what they measured was *actually* motivation. This is referred to as the 'validity' of a scale [18].

In regards to this issue, recent research has shown that in psychology there is room for improvement in ensuring the validity of measurements [7]. One first step identified in this regard is the transparency of reporting measurements [22], which must be given to judge whether a measurement is valid within the context it is employed in. Reviewers and readers of a study require specific information about the construct a questionnaire is measuring (e.g. 'Affective State') or how a scale was administered to participants (e.g. 'The Positive and Negative Affect Schedule consists of 10-item scales (1=not at all; 5=very much) resulting in two sum scores, one for positive affect and one for negative affect.'). Without this information it becomes impossible to evaluate whether the measurement is valid in its given context [3]. This extends to the data and the subsequently drawn conclusions being questionable in their validity. Meaning, an issue with validity in measurement will contaminate all conclusions drawn from the data collected with those measurements [35]. It is for this reason researchers, who are invested in the validity of measurements, have called for everyone to engage more strongly with the measurements they use for their studies and the literature surrounding them [8, 47].

Within the field of PX research, there are further specific concerns which have been raised. Specifically, certain commonly used questionnaires, such as the Game Experience Questionnaire (GEQ) [28, 42], have been questioned in whether or not they are a valid measurement of the construct they propose to measure [11, 29, 31, 38].

It is important to acknowledge that the measurement of subjective experience through self-reported survey scales is only one of several research methods used at CHI PLAY. For example, and this is by no means exhaustive, player behavior and experience is also explored using qualitative methods such as grounded theory [9] and ethnomethodology-informed ethnography [50] or with AI agents in simulations [43]. Indeed, when submitting a paper to the CHI PLAY conference, authors are able to choose from many different contributions beyond research which involves the quantitative measurement of constructs [2]. These contributions are directly derived from the many forms of research which has been published in HCI [57]. This highlights that CHI PLAY is a diverse and multidisciplinary community (see also Carter et al.'s Paradigms of Games Research in HCI [12]) and so are the scientific methods used to generate knowledge.

Due to this multidisciplinary field of research, transparency has become an important topic in many different forms. Indeed, transparency has been investigated in regards to sharing data, analytic methods or using preregistration to record possible changes to the research design over time [14]. Further, transparency of materials such as source codes [19] and research artifacts have been investigated [55]. However, none of these works has yet reviewed transparency of measurement reporting specifically, as such, further work in regards to transparency of research is needed.

Additionally, best practices on measurement reporting are not commonly shared within the research community. This uncertainty of the measurement quality in PX research is in opposition with the recently expressed aim of the CHI PLAY. Namely the aim that CHI PLAY papers be included in meta-analytic research [36]. If one can not be certain of the validity of a measurement, generalizing the construct measured over multiple studies would be a fruitless endeavor [20]. Therefore, it is crucial that the validity of the measures is investigated and that the measures are reported transparently. This does not only concern self-reported survey scales but also objective measurements. However, reporting and establishing the validity of scales is arguably more challenging than, for instance, of time played or number of levels completed.

The contribution of this systematic literature review is therefore threefold: First, the review of literature provides an accessible discussion of the validity of measurements and its importance and elaborates how this relates to general concerns in regards to the generalizability of studies and consequently their inclusion in meta-analyses. Second, we examine the reporting of survey scales at CHI PLAY 2020. This will provide a first understanding of the state of transparency in the reporting of survey scales within PX research. Third, we provide a model of the measurement selection process for PX researchers to employ. This should aid in achieving transparent measurement reporting, therefore improving the validity and generalizability of findings.

2 RELATED WORK

2.1 Survey scales in research

Survey scales and other self-reported surveys have often experienced a negative stigma due to the methodological biases which they can potentially introduce [15]. For example, reviewers in organizational psychology often question the legitimacy of self-reported questionnaires in favor of more objective data [13]. However, evidence suggests that frequently discussed biases, such as social desirability, do not have strong, consistent effects [56].

2.1.1 Validity concerns of survey scales. The concerns over self-reported questionnaires can, despite this, still be valid. In specific, user experience (UX) researchers have expressed doubts in regard to whether measurements can accurately capture the experience of users, as empirical data is meaningless without theory [32]. In the broader context of empirical research, substantial method effects threaten the validity of results drawn from measurements. Validity is commonly understood as the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores [5]. DeVellis [18] defines three forms in which these measurements need to prove their validity. First, content validity, which refers to the fact that a scale's content should reflect the conceptual definition applicable to that scale. Second, criterion validity, also sometimes referred to as 'predictive validity', is a theory-neutral form of validity. Criterion validity assesses whether the empirical relationship between a scale or item and a criterion, or 'gold-standard' is strong [18]. Third, construct validity, which places the measurement in relation to other measurements, ensuring that the intended construct is being measured and not an alternative one [15, 40]. In line with previous research, we define constructs as theoretical concepts aimed at organizing and making sense of our environment [40]. These constructs are not directly observable or measurable [26]. Construct validity can be evidenced using different statistical methods, however it is impossible to definitively prove construct validity [40]. Despite this, methods exist to provide evidence of construct validity. These methods are generally referred to as the validation of a survey scale. Included in the validation can be an analysis of factorial structure, appropriate reliability measures, mean differences between relevant groups (known-groups validation) or evidencing divergent validity by measuring a construct that is supposed to be distinct [15]. Often this information is present in studies which validate a particular questionnaire. In subsequent studies, which employ the scale, this information is made available to reviewers through references. However, the epistemological discussion surrounding validity, while having impactful consequences for our research, is much deeper than we can discuss in this paper.

Indeed, this paper is instead a first step towards transparency in measurement reporting. We argue that beyond providing evidence of construct validity through reference, researchers have an obligation to provide enough information to judge the specific use of a questionnaire in their study. The content of this information will be dependent on the study. However, information to answer the necessary questions a reviewer might have about the appropriateness of a measurement should be present in all studies which employ questionnaires.

2.2 General concerns when reporting measurements

As explained above, the constructs we measure and the measurement we chose for these constructs are a fundamental challenge of every empirical study. In their paper Vornhagen et al. [54] gave a broad idea as to how theoretical variables should be reported in studies. They explain these constructs must be precisely defined, the measurement procedure must be elaborated on and the selection of measurements must be justified. In this section we will review psychological literature on measurements and the reporting of them to explain why this information and more is crucial to ensure transparency of measurement reporting.

2.3 Defining constructs

In line with Vornhagen et al. [54] we discuss why precisely defining a construct before its measurement is necessary for interpretable and generalizable findings. Referencing DeVellis [18] conceptualization of content validity, researchers cannot be certain that the content, or items, of a scale is appropriate for their chosen construct if they do not have a clear definition of their construct. Being uncertain about what it is that is being measured, makes it impossible to precisely measure this construct. MacKenzie [35] elaborates how inadequately specifying the conceptual meaning of the study's focal constructs will inherently lead to undermined construct validity. Undermined construct validity further contaminates internal validity and statistical conclusion validity. Meaning that researchers cannot be certain of the results, or the implications drawn from them, which they generate using scales with undermined construct validity. This presents a serious threat to the generalizability of the findings of studies with ill-defined constructs. It further hinders the development of theory, as theory is built upon the building blocks of well-defined constructs [48].

This is an even larger issue, as it has been noted in psychology that certain constructs can be studied for decades and still remain ill-defined [3]. Antonakis [3] details how these ill-defined constructs hinder the advancement of research, as one can never be certain if a study accurately captured the construct they meant to capture. Often, to bridge the troubled waters of ill-defined constructs, researchers will use measurements to theorize about their constructs, rather than using theory to construct their measurements.

2.3.1 Specifying constructs in theory. Constructs should be specified within a certain theory [24]. If this theory is lacking when measuring constructs, it has further implications. One large issue with relying singularly on statistical processes to assess validity of a measurement is that they can be used while disregarding theory [21] and rendered meaningless because of it [37]. Indeed, this leads to theory being not prospectively described and operationalized. Meaning, theories will lose their predictive value, leading to difficulties when hypothesizing and designing studies.

In their satirical research paper Satchell et al. [45] validate a scale measuring 'Offline Friend Addiction'. They do so by defining the construct of 'Addiction' by the outcomes of the measurement, following the methods of other commonly used social media addiction scales. This leads to a conceptualization of addiction as anything in which people passionately engage in employing repetitive behaviors, whether or not these behaviors are actually pathological. This shows the problematic nature of defining constructs by the scales that measure them. This lack of theory to construct statistically valid questionnaires has been further examined in a study by Maul [37]. Maul finds that meaningless scales using nonsensical terms can be evidenced as valid using methods such as factor analysis, as participants generally answer consistently as long as the same answer format is given. This consistency is not due to actual construct validity, but will satisfy the same statistical criteria as a meaningful survey scale being validated in the same manner. Therefore, statistical processes of validation of survey scales are only meaningful when the scales are built on solid theory supporting them.

2.3.2 Poor Definitions and their Implications. Poor definitions of constructs can be recognized through the following qualities [3]: Constructs should never be defined by their outcomes or ascendents. Further, if a construct is frequently equated to other constructs, it could indicate a lack of divergent validity. Finally extremely broad definitions are also not useful for their precise operationalization. However, often constructs are not defined by the researchers in the first place. What researchers ought to do is to combine the definitions of previous literature to create their own [35]. In contrast to this, many psychological researchers instead provide an overview of previous definitions without defining their own [3]. This makes it impossible for reviewers to judge the accuracy of the selected measurement, as a concept such as 'engagement' could be understood by the authors in multiple different ways and consequently different measurements would be more or less accurate. Ekkekakis and Russell [21] explain that researchers have become desensitized to fundamental issues of imprecise terminology and blurred conceptual distinctions. Due to their frequency, researchers can gloss over such inconsistency and imprecision when reviewing papers. In many disciplines the clear distinction of terms is still a rather young phenomenon. As such it is important to inform researchers who attempt to precisely measure a certain construct of best practice standards when defining constructs [21].

2.4 Operationalization of constructs

Constructs are theoretical concepts, which are not directly observable [26]. This means a translation from the theoretical concept into a measurable object has to be created. This translation of something not observable into something observable, something not measurable into something measurable, is referred to as the operationalization of a construct [49]. The idea of operationalism (e.g., [53]) behind the practice of operationalization is an important discussion, but out of scope for this paper. Indeed, we will instead focus on practical concerns when reporting the operationalization of constructs.

How a construct is operationalized can introduce methodological variability. Meaning that difference in measurement presentation or administration can influence the results gathered from a survey scale [22]. Indeed, issues such as minor wording changes [34], or the amount of points on a Likert scale [30] can influence the answering behavior from participants in studies.

Further ambiguity around how a measure was precisely operationalized contributes to a lack of continuity in research [22]. This means that researchers who aim to measure precisely the same construct in the same way (e.g., for a close replication), would be unable to and would have to guess, introducing possible differences in results due to methodological variability. Further, should researchers not report exactly how the measurement was operationalized and administered, it is impossible for reviewers and meta-researchers to know how a construct was measured.

2.5 Measurement selection

When researchers make the decision to employ a survey scale, they are immediately confronted with yet another decision. Which questionnaire of the many should one use to measure their constructs? When one looks into previous research one can find a multitude of questionnaires measuring common constructs such as enjoyment, so which of those is the most appropriate one? Or the opposite could be true; after looking for a while no questionnaire that fits the construct the researcher has imagined appears. According to Ekkekakis and Russell [21] researchers sometimes like to pretend no such thought process has occurred within them. In those cases it can appear as if a measure appears out of the blue in the measure section with no justification for its presence given. Even when justifications for measurements are given they are often superficial. Researchers refer to how commonly a measurement is employed or its age. Neither of these aspects are necessarily indicators of the appropriateness or quality of a questionnaire. Ekkekakis and Russell also describe that researchers continue to use questionnaires which have received valid critiques.

This idea of negligence towards measurement selection is supported by Pedhazur and Pedhazur Schmelkin [40] who note that researchers often treat measurements 'mindless'. Measurements appear to be used simply because they 'exist', with researchers investing naive faith into them.

As discussed before, Ekkekakis and Russel further elaborate how when researchers select a scale, they should understand and agree with the underlying theory of this scale [21]. This is especially important in light of the meaninglessness of statistical validation without a theoretical conceptualization of a construct [37]. This leads to researchers not only using a scale, but also using and supporting the theory which was used to build it, however, the majority of papers do not explicitly and meaningfully discuss this theory they have chosen in tandem with their scale [51].

A further concern lies in a common reason of measurement selection, namely its seniority [21]. Indeed, researchers should not simply 'grandfather in' scales which have been employed before solely on the basis of their previous use. Pedhazur and Pedhazur Schmelkin note that certain scales, such as the F(ascism) scale by Sanford et al. [44] stemming from the 1950s would not be appropriate for use in its original state, as the context in which the items are posed has changed drastically since then [40]. As such, even if scales have been employed previously by other researchers, it is important to note for what reason their use is appropriate in the current study.

Therefore, there is a need for a rigorous selection process for measurements which is reported for each study, so that reviewers can judge the validity of any chosen measurement.

2.6 Self-developed measures

So far we have discussed validated questionnaires and their potential issues with validity. However, not for every study conducted in PX has an appropriate measurement already been developed and validated. This is a situation that occurs commonly in HCI research [6, 41]. In these cases researchers might choose to employ self-developed measures instead. The self-developed measures we refer to function similarly to validated questionnaires [27]. While the use of these self-developed measures is not inherently problematic, often no assessment of the quality criteria, including validity, is made.

As there is no validation study to cite, often no information in regards to construct validity is given. This can make it impossible to judge whether a measurement actually assessed the construct it was intended to. When using self-developed measures, researchers need to take additional care to provide information to make it transparent to reviewers whether their measurement was valid [22]. In contrast, validated questionnaires (should) provide the information to judge their construct validity through citation. Further, a validated questionnaire (should) be worded the same each time it is employed. This greatly increases the ease of comparison between studies and especially the generalizability of findings [27].

Therefore, it is advised to always use validated questionnaires whenever possible. This means that constructs which are measured in the field of PX research using a validated questionnaire should not be measured by self-developed measures in a different study without justification [33].

3 RESEARCH GOALS

In line with the previously discussed literature this systematic literature review aims to contribute to the quality of future PX research in the following ways:

We aim to provide a systematic review of the current state of transparency of measurement reporting in PX research. Combining these findings we aim to devise a model of a measurement selection process researchers can employ as a first step towards transparency in measurement reporting in PX.

This study does not aim to identify any questionable measurement practices. We only examine the transparency of measurement reporting necessary to be able to identify potential questionable measurement practices. Further, we aim to not single out specific studies and their authors, but only to provide examples of transparent or intransparent measurement reporting.

4 METHODS

We conducted a systematic literature review, which we report on following the PRISMA 2020 checklist [39]. We investigate the current state of transparency of measurement reporting at CHI PLAY 2020. The PRISMA Flowchart detailing all steps of the review can be found on https://osf.io/xbjw6/.

4.1 Eligibility and Information source

For this literature analysis, we collected all research articles, or full papers, published at ACM CHI PLAY 2020 (n = 48). Excluded from the final sample were all non-full text submissions, this was done to ensure that all papers reviewed are in themselves complete and should include all information in regards to their measurements.

4.1.1 Search strategy and Selection process. We used the ACM Digital Library[1] to download the full proceedings of CHI PLAY 2020. All 48 papers were coded for their structure and then screened for the inclusion of self-reported measures. When papers included self-reported measures they were further analyzed according to the following code book.

4.2 Potential Bias

No risk of bias due to reporting biases (e.g., publication bias) should be present in this study, as our goal is to assess transparent measurement reporting in only published full papers. This is done as to not implicate the research field with intransparency of measurement reporting in papers which did not pass peer-review or are not completed. We acknowledge there is a potential bias as the proceedings of CHI PLAY 2020 could differ in terms of transparency from other years. However,

[1] https://dl.acm.org/conference/chi-play/proceedings

the recency of these proceedings should provide us with the current transparency of measurement reporting as a fair basis to assess potential pitfalls.

4.3 Code book

Using the questions to promote transparency of measurement reporting proposed in Flake and Fried, we developed an initial code book. The early categories followed six questions proposed by Flake and Fried closely [22]. However, we excluded the investigation into the quantification of the measurements, as the focus of this paper lies in transparency of study design and measurement selection. As such there were codes in the category of 'Construct Definition', 'Construct operationalization', 'Measurement Selection', 'Modification of Measurements' and 'Self-Development of Measurements'. Post-hoc statistical transformations of the scale were not coded for. We developed the code book further in the following way:

- We structured the coding by isolating each construct measured by a paper and investigating the level of transparency of measurement reporting for each one.
- For our purposes, we decided to record a measurement as self-developed when the authors did not cite a source of that measurement.
- We further recorded which scale was cited, in case of a supposed pre-validated questionnaire, to examine whether the cited papers offer evidence of validation.
- As will be elaborated in the results section, many papers did not offer easily accessible information in form of a 'Methods' or 'Measurement' section, as such we included codes which identified the presence of these sections and what other names are given to similar sections.
- We added a text code for the descriptions of modifications made to questionnaires and scales. This was done to record what kind of modifications were reported.
- In line with Flake and Fried we implemented a code to assess whether modifications of questionnaires occurred before or after data collection. However, we found that for some modifications it was implicitly known whether modifications occurred beforehand, we added a secondary code to record these instances.

An example of a code would be:

> Construct defined (0/1): Does the construct have a definition in the paper. Quality is not important, any statement of 'x is defined/means/is' should be coded as 1. 0 should be coded if there is no definition, no matter how common the construct seems, e.g. 'enjoyment'.

The full code book can be found on https://osf.io/jxbg2/

4.4 Procedure

In order to test the quality of the initial code book a randomly selected subset of 5 papers (10.42%) was coded by the first author to revise the code book accordingly. We introduced a full-stop into the code book for studies who did not measure a construct using survey scales (50% or 24 of 48). After the code book was finalized, the full sample was coded independently, with the first author coding 30 papers, the second author coding 15 and the third author also coding 15. This created an overlap of 12 papers coded by two researchers independently with which we calculated inter-rater agreement, this was 86.61% for all numerical codes. In a secondary step we discussed mismatched codes until consensus was reached and finalized the coding for the interpretation of results. After the initial coding was complete, we wanted to assess whether the given citations for each scale contained evidence of validity, as such the first and third author analyzed these secondary sources in regards to whether they examined the factorial structure of the questionnaire.

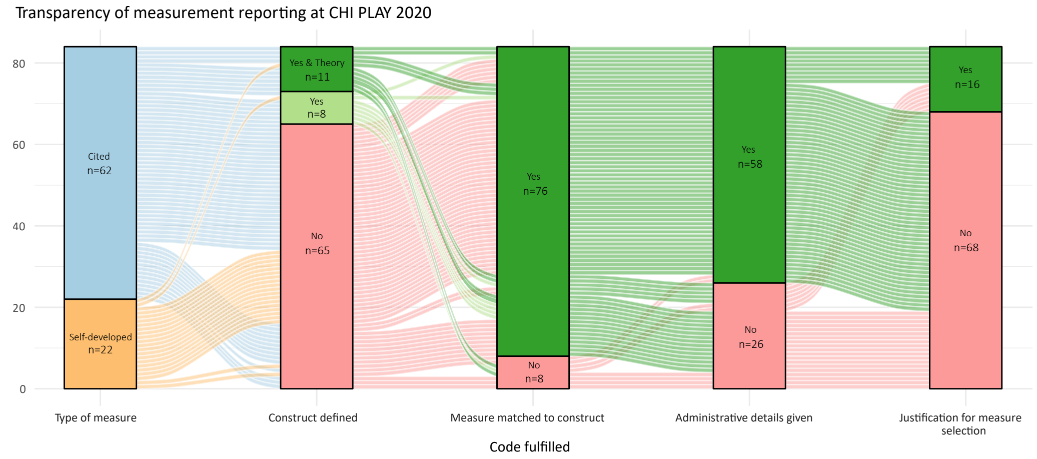


Fig. 1. Number of codes fulfilled for all instances of measurement. N = 84.

5 RESULTS

We present findings in the general order of our code book. We report the results from 24 studies in full, no further eligibility criteria was applied. No data was converted or transformed, as we only counted occurrences of transparency and report on these counts and their percentages. We emphasize again that this review attempts to assess the transparency of measurement reporting in current research, not find questionable measurement practices or lack of validity of measurements. Our full sample was comprised of 48 full papers, of these, n=24 or 50.00% reported measuring at least one construct using some form of self-reported survey instrument.

5.1 Structure of papers

To find information about measurements quickly and efficiently, dedicated sections to both the methods and the measurements used aid readers [4]. This can allow researchers to quickly scan papers for this information. We therefore examined whether the paper had sections titled 'methods', 'methodology' or 'method' and further a section titled 'measurements'. Were no such sections present, we instead recorded the presence of a section describing the methods or one describing the measurements and further recorded their names. We found that in the total sample of 48 papers, 23 of 48 papers (47.92%) included a section titles method, methods or methodology. Further 22 of 48 papers (45.83%) included a section which was similar to methods, but it was named differently, examples of these names include 'Procedure' or 'Experiment N'. Only 3 of 48 (6.25%) had no section to detail their research methods. We also recorded whether a specific measures subsection was present in the 24 papers in which at least one occurrence of measurement was reported, we found that 12 of 24 papers (50.00%) which measured at least one construct, included a measures subsection. On the other hand, 11 of 24 papers (45.83%) included a section which included the information about measures, but named differently, with names such as 'Procedure' or 'Tasks'. As discussed above, we recorded these instances as we find a unified language and structure of a research field an important step towards meta-research.

5.2 Constructs and Measurements

In total, we recorded 84 instances of measurement, of those 62 employed cited measurements and 22 employed self-developed measurements (see the first bar 'Type of measure' in Figure 1). We found 67 different constructs. Enjoyment was the most popular construct, with it being measured in six different instances. Most of the constructs were unique, only being measured once throughout the 24 papers (n = 60). This breadth was further reflected in the recorded measurements, we found 41 different cited measurements used, while 22 measurements were self-developed by the authors of the papers.

5.3 Defining constructs

As described previously, the first step to measure a construct is to define it. Unless a researcher clearly states how a construct is to be understood in the context of their study, it becomes impossible to assess whether their measurement was appropriate. As defining a construct by its measurement is not appropriate either [3, 45], researchers need to state a clear definition of how the researched construct is to be understood. Simply matching a construct to a measurement without a proper definition does not provide enough information to readers and reviewers. In total, we found 19 of 84 constructs measured (22.62%) were defined. 11 of 84 constructs were specified within a theory (13.1%). We found that 12 of 24 papers (50%) defined at least one construct, while only 2 of 24 papers (8.33%) defined all of their constructs measured. Examples of definitions of constructs are as follow:

> "Self-presence represents the connection between players and their avatars, which guides people to choose what to present via their avatars" (N21).

or for the construct "Creativity":

> "Creativity is expressed in different forms depending on the nature of the task and the medium of creative expression figural creativity (e.g., drawing, painting, sculpting) and verbal creativity (writing, storytelling, composition, discourse) [22]. In this work, we focus on figural creativity that is illustrated through a drawing game." (N5).

For cited measurements we found that in 17 of 62 instances (27.42%) authors defined the constructs they measured.

We also specifically recorded the percentage of self-developed measures being defined and found 2 of 22 self-developed measures (9.09%) were defined. For example, one definition of the construct 'Game Atmosphere' given was:

> "The term 'game atmosphere' is used to describe a subtle but important, intangible, generally aesthetic quality in games that leads to emotional immersion." (N36).

or the construct "Purchase Intention", which was defined as:

> "[...] the desirability of owning the game [...]" (N1).

5.3.1 *Specifying Construct within Theory.*

For a complete understanding of a construct, the definition needs to be specified within a certain theoretical framework [24]. For this we collected how many authors provided information in regards to which theory their construct is specified in. A construct was considered as derived from theory, if authors gave any reference to the theory in which their definition of a construct is to be understood. We found that 11 of 19 defined constructs (57.9%), were specified within a theory. One example of being transparent in regards to which theory informs their construct definition was as follows:

> “Based on Goffman’s metaphor of theatrical performance[14], the notion of self-presentation online highlights how identity is portrayed and experienced as a combination of conscious personal choices and specific technological affordances of online social spaces” (N21).

A visualization of the rate of definitions given to measured construct can be found in the second bar (’Construct defined’) in Figure 1.

5.4 Operationalization of constructs

Two codes were used to record the transparency in regards to the operationalization of constructs. How precisely a construct was operationalized is important to know for the reviewers who need to judge whether a specific measurement is appropriate for the construct. But further, it is vital for potential replications of research, as without it the precise method would be not replicable.

5.4.1 Matching measurements to construct. The first code recorded whether the authors matched their cited measurements to their constructs. We found this to be consistently the case, with 76 of 84 constructs matched to the cited measurement used (90.48%). An example of a such a matching is:

> “Participants’ enjoyment was measured using the enjoyment scale items from the Intrinsic Motivation Inventory” (N19).

When self-developed measurements were reported this was done especially consistently with 21 of 22 (95.45%) matching them to the constructs they measured. However when examining cited measurements this was still consistently done in 55 of 62 instances (88.71%). This code is represented in the third bar (’Measure matched to construct’) in Figure 1.

5.4.2 Administration details. The second code was concerned with whether authors would be transparent about the administration of a measurement. Explanations of administration could include such things as the amount of points on a Likert-type scale or descriptions of the digital environment in which participants were present when filling out the questionnaire. Authors of CHI PLAY also relatively consistently provided this information in their papers. As such, 58 of 84 measurements (69.05%) included information about administration procedure. A typical description of this administrative procedure was as follows:

> “Each subscale includes 4 items, ranked on a four-point Likert scale from “Not at all” to “Completely”” (N18).

Again, this was especially consistent for self-developed measures with administrative details being given in 21 of 22 instances (95.45%). This was done less consistently for cited measurements with 37 of 62 measurements providing this information (59.68%). This code is represented in the fourth bar (’Administration details given’) in Figure 1.

5.5 Justification for measurement selection

As discussed in the related work section, authors should provide a justification for the specific measurement they cited. Justifications could include statements about widespread use or about the validation of a questionnaire. We emphasize that this code was merely about the presence of such a justification and we did not make any judgement about the validity of these claims. We found that authors often neglected to explain why they selected a measurement, with 16 of 84 total measurement choices (19.05%) being justified by researchers. When citing a measurement, researchers justified their selection 16 of 62 the time (25.81%). Justifications ranged from very simple statements such as:

> “a well-validated measure of intrinsic motivation for a task” (N17),

to more in-depth explanations:

> “We opted for this univariate scale because it is less extensive as other scales such as the NASA Task Load Index [25], which might have become laborious given that it needs to be filled out for each task. In addition, previous work [54] indicates that it is reliable and sensitive to small differences in task complexity” (N3).

For self-developed measurements in 0 of 22 instances of measurements (0.0%) authors justified why they created their own measurement. This code is represented in the fifth bar (’Justification for measure selection’) in Figure 1.

5.6 Modification of measurements

We recorded all described instances in which authors made modifications to their measurement. However, it is possible that authors have neglected to describe minor modifications such as changing the phrasing of ‘this activity’ to ‘this game’ in questionnaires such as the IMI. We found that 24 of 62 cited measurements (38.71%) were reported as modified. When modification was reported 9 of 24 of modified measurements (37.50%) gave a justification for the modification, while 21 of 24 modified measurements (87.50%) gave details or examples for the modification. An example of such a description of modification is as follows:

> "Furthermore, we adopted the questionnaire to transfer the individual items to our game setting. For example, the original version “I could readily tell when my partner was listening to me.” was replaced with “I could readily tell when the game entities were turning their attention to me."" (N12).

Further in 0 of 24 cases (0.0%) where modification occurred, authors reported whether this modification was done prior or after data collection, however in 4 of 24 (16.67%) it was implicitly clear whether modification had occurred before or after data collection.

5.7 Self-developed measurements

As described above, self-developed measures require more information than validated surveys, as the validation study of a survey provides evidence for construct validity. We found that 12 of 24 of papers (50.00%) included self-developed measurements. Of those self-developed measures 17 of 22 (77.27%) provided details or explanations of their measurement. An example of such a description of a self-developed measure is:

> "7-point Likert scale questions regarding the preferred level of realism of the agent "I would have been more interested in customizing the agent if it was more cartoon-like/realistic.""(N19).

Authors provided evidence of construct validity for 4.54% (1 of 22) of the self-developed measures, in this case authors calculated an exploratory factor analysis. The rate of self-developed measures to cited measures is represented in the first bar (’Type of measure’) in Figure 1.

5.8 Validity evidence for cited measurements

So far we have used the term ’cited measurements’ in the results section, rather than ’validated measurements’, due it being unknown when only coding the original papers whether the citations provided had any evidence for construct validity as one would assume. In order to assess whether the citations provided evidence of construct validity, we coded the cited papers for all measured constructs. Meaning we searched the 41 cited papers for evidence such as exploratory or confirmatory factor analysis. We found 19 of 41 citations (46.34%) did not provide any evidence of construct validity. In a tertiary step we attempted to find any papers, not those which were cited by the authors of CHI PLAY, which provided construct validity evidence for the employed survey scales. In this manner we could find that 30 of 41 the survey scales (73.17%) cited by authors offer evidence of validity at one point or another. However, the scale which was cited and used could differ from the scale which showed evidence of validity. As such, it is unclear to us whether the construct validity reported in those tertiary papers was applicable to the version of the scale employed by the researchers at CHI PLAY 2020.

6 DISCUSSION

The first aim of this study is to provide an overview of the current state of transparency of measurement reporting. Our findings show both strengths and weaknesses in regards to transparency. As such we present a discussion of the findings from our systematic literature review of the CHI PLAY 2020 literature. We note that our data only includes papers published at CHI PLAY 2020, as such these findings do not lend themselves to implicate all of PX research. Instead, we describe current trends found in our data and their broader implications. Further, this research needs to be contextualized in the multidisciplinary field of PX research. This paper does not aim to prescribe the quantitative measurement of constructs as a one-size fits all solution to our field. For example, the contribution of artefact research is important without the empirical data of quantitative measurements. Instead, our aim is to amplify the importance of transparency in measurement reporting and validity of measurements to draw robust conclusions from these empirical findings. With the concerns raised by this paper in mind, researchers might find that their research is not significantly improved by the inclusion of measurements. Therefore, their artefact based research should not include the quantitative measurements of constructs [23, 57]. This way, researchers not primarily concerned with the outcomes measured with survey-scales, do not have to implicate their findings with issues of measurement validity.

6.1 Findings from systematic literature review

6.1.1 Defining constructs. In our assessment of current transparency of measurement reporting we found that intransparency often begins at the very basic question of *what* is being measured. Less than a quarter (19 of 84 or 22.62%) of all constructs which were measured were defined at all. As shown in the method section, our code did not distinguish between correct, incorrect, complex or simple definitions. A simple statement of what a construct is or means was coded as this construct being defined, yet such statements were absent in most instances of measurement. This means that reviewers will be unable to judge whether a chosen measurement is appropriate for the chosen construct as understood by the authors. As described in the related work, being uncertain or unclear about what is being measured will undermine all results originating from that measurement in their validity [3, 21, 35, 45].

But this lack of definition also extends to the broader context of research. While definitions can not be proven, they need to be consensually considered useful by a research community to make research comparable and cumulative in meta-research [46]. However, such a consensus can only begin to form when researchers consistently invest effort into conceptualizing the constructs they study. PX research is at risk of running into similar issues as researchers in other disciplines in which they attempt to measure complex psychological processes. For example health-behavioral [21] or organizational sciences [3]. Namely a lack of a cohesive understanding of central concepts and how to distinguish them from other similar but not synonymous constructs.

The absence of consistent definitions can further lead to a void of theory due to brittle conceptual building blocks. However, the theory behind constructs is even more rarely reported within our sample of papers. In regards to specifying constructs measured in their respective theory it only occurred in 11 of 84 instances of measurements (13.1%). This is also problematic in multiple regards.

In the absence of theory the results of statistical validation procedures such as confirmatory factor analysis become essentially meaningless [21, 37]. This, again, undermines construct validity of a measurement and subsequently the results drawn from it. However, these issues are not constrained to the studies in which they occur, but, again, extend into the entire field of PX research and beyond. Especially with the expressed aim to be included in meta-analytic research [36], it is important to build solid theories that can be generalized over all of PX. However, if the constructs are not specified in a specific theory conceptually, studying them and reporting on their effects will not aid in theory-building [24].

6.1.2 Operationalization of constructs. We found that it was often transparent *how* a construct was measured. Meaning, the operationalization of constructs was often easy to discern both for cited and self-developed measurements. When using a measurements, researchers would most often (76 of 84 or 90.48%) match it to the construct they measured. This way reviewers can discern whether they find the selected measurement appropriate for the construct given. Further, details relating to administration such as the points of the Likert-type scale employed were often (58 of 84 or 69.05%) given. This shows that researchers at CHI PLAY are aware of the requirements of providing such information. This also allows for reviewers to assess whether the measurements used were implemented appropriately.

When researchers developed their own measurements, they gave information of the actual content of their scale. Meaning they provided examples of the items they used or gave explanations in regards to them, making it possible to understand how they measured their proposed construct (17 of 22 or 77.27%).

This is reflected in the coding of the modifications done to cited measurements that were also commonly described by the researchers. This is important, as statistical processes that evidence construct validity can be made non-applicable by severe modifications [22]. Describing the kind of modification that was done for a measurement allows for reviewers to judge the appropriateness and whether the original validation still applies or researchers need to validate their own modified version.

6.1.3 Justification for measurement selection. A second issue is the intransparency in regards to the process of measurement selection, or *why* the chosen construct was measured in this way. A fifth (16 of 84 or 19.05%) of all measurements stated any kind of reason why they selected their specific measurement. For example, despite recent research questioning its validity [11, 29, 31, 38], the GEQ [28, 42] was used in our sample with no justification as to why (N28). This can indicate a lack of a rigorous measurement selection process in which authors make informed decisions as to which measurement is best suited for their chosen construct.

6.1.4 Modification of measurements. In regards to modifications, we also found that no researcher explicitly stated whether a modification occurred before or after data collection. In case of changed wordings, this can be considered negligible, as implicitly it is known that modification had to occur before data collection to impact the findings, but for other modifications it is imperative to know in what part of the research process they happened. Dropping individual items or dimensions of scales during analysis could result in accidental p-hacking [25]. Similarly, less than half of the time (9 of 24 or 37.50%) researchers reported why they had modified the measurements they used. Any modifications and justifications as to why they had to occur would also be part of a transparently reported measurement selection process. However, if modifications occur before data collection, are justified and pose no threat to construct validity, they can be beneficial to the study.

6.1.5 Self-developed measures. In comparison to similar research in UX studies, we found a similar amount of studies employed self-developed measures. In their study Pettersson et al. find that 53%

of papers used at least one self-developed measure, while in our sample we found a rate of 50% [41]. When examining the use of self-developed measurements, further issues arrived. We found that some researchers developed their own measurements for constructs that other researchers at CHI PLAY 2020 employed a validated questionnaire for (e.g., 'Enjoyment'), while other times a self-developed questionnaire was used for constructs for which validated survey scales exist (e.g., 'Satisfaction'). As no researcher provided reasoning for why they developed their own measurement, reviewers and meta-research can only guess why they chose to create their own measurement. The reasonings could be issues of implementation (e.g. length of a validated survey scale) or theoretical (e.g. different understandings of how a construct should be defined), but it remains unclear to readers of the paper. Justification for why a self-developed measurement was employed should always be given [22], but are especially important when validated measurements already exist for the same construct [27]. Further, in only one instance (1 of 22 or 4.54%) did authors provide an investigation into the factorial structure of a self-developed measure. Meaning for only one self-developed measurement evidence of construct validity was reported. This directly impacts how certain we can be of the implications drawn from the results of the measurement, as reviewers and meta-researchers can not evaluate whether the measurement measured what it set out to measure [18].

6.2 Further implications of intransparency

As this study investigates transparency, a lack thereof leads to a number of possible implications depending on the information which was not reported by researchers. We aim to report on the most important possible implications of potential issues in this section.

6.2.1 Jingle-Jangle Fallacy. Intransparency in regards to both definitions and selection can lead to the so-called 'jingle-jangle fallacies' [22]. The jingle fallacy refers to the issue in which researchers assume measurements with similar names measure the same construct, while that is not actually true. The jangle fallacy refers to the inverse, two measures measuring supposedly different constructs because their names are different, while in reality they measure the same construct. We find that there is a possibility for jingle-jangle fallacies to occur as certain constructs, such as 'Enjoyment', are measured with different scales without justification or definition. With a lack of explicit definitions and justification of measurement selection, it is difficult for reviewers to know if researchers who measured enjoyment with the Intrinsic Motivation Inventory (IMI) [17] conceptualize and theorize it in the same way as those who measure it with the Player Experience Inventory (PXI) [1]. As such, researchers also need to start defining their constructs and justifying their selection of measurement to address potential jingle-jangle fallacies.

6.2.2 Confusing and Incorrect Citations. Another confusing issue arrived when recording the citations of questionnaires. We found certain authors cited studies that offered no evidence of construct validity for the questionnaire they used (46.34% or 19 of 41). Alternatively, some citations were simply incorrect, such as in the case of the GEQ, which has a history of confusing citations [31]. We also found researchers citing different studies for measurements of the same name. This can lead to confusion as it is unclear if there are multiple versions of the scale validated under the same name. All of these issues regarding citation made it near impossible to assess in certain cases whether a survey scale that was employed had been validated previously or whether evidence of construct validity for this survey scale was simply lacking. This issue will be similar for reviewers who try to judge the validity of a scale. A reviewer could assume that a citation of a scale means that this scale has been validated or else the researcher would not employ it without providing their own validation procedures. Therefore, we urge researchers at CHI PLAY to be careful with citations and to not report a survey scale which lacks evidence for construct validity in the same manner as one which does offer such evidence of validity.

6.3 Measurement Selection Model

The second research goal was to formulate a prescriptive model of the measurement selection process (see Figure 2) in line with research models devised by Guest and Martin [24] or Ekkekakis and Russell [21]. We believe that researchers employing our model will promote transparency in regards to aspects of measurements which are currently seldom reported systematically on in CHI PLAY literature, such as definitions of constructs and justifications for measurements. Improved transparency of measurement reporting in turn will aid reviewers and meta-researchers to fully grasp the entire design of a study to judge its appropriateness or inclusion in any given meta-analysis.

6.3.1 Defining constructs: What are you measuring? In line with the discussed literature in this study [3, 35], we understand measurement as beginning with the conceptual and theoretical definition of a construct, rather than the operationalization of that construct. While in our literature research we did not investigate the quality of definitions given, there are still certain aspects that make something a *good* definition.

The choice of construct should be given a lot of weight. Researchers need to be certain that this construct was the most appropriate for the given context. As such researchers should also remain specific and not use near synonymous terms within the same study [21]. Similar terms such as 'Immersion' or 'Engagement' should therefore not be used interchangeably in the same study.

Once a specific construct is decided on, the precise definition should be reported in the study. Good definitions should fulfill a number of standards. Firstly, good definitions should specify the constructs conceptual theme and be unambiguous. Meaning they need to be clearly distinguished from related constructs. Further, if the construct is multidimensional, the relationship between the dimensions need to be explained [35]. It is also important for researchers to take ownership over the definition they use in their study. Researchers should give a statement how *they* defined the construct in *their* study, not simply how it has been defined previously [3].

This definition further needs to be specified within a theory and not simply stand for itself [24]. Understanding which theoretical underpinnings inform a construct is necessary to fully define a construct.

6.3.2 Operationalization of constructs: How are you measuring? Once a construct is defined both conceptually and theoretically, researchers can move on to considerations how exactly they will measure this construct. Should a validated questionnaire exist with the same conceptual and theoretical background as the construct defined, this should always be chosen over developing your own survey scale [27]. As changing numbers of Likert scales can change the results collected from experiments, researchers should use the validated format [16]. Further modifications, including the wording of questions, translation or time specification in recall periods, can also impact their validity [30]. As such they must be carefully considered and fully reported.

Should no appropriate validated questionnaire exist, researchers can develop their own measurement. These measurements must also be in line with the theoretical considerations of the construct.

6.3.3 Justification of measurement selection: Why are you measuring this way? All of the previous considerations combined should give the first part of the reasoning as to why a certain measure was selected or developed. In addition to considerations in regards to definition, theory, administration and operationalization as described above, a comprehensive report on validity must be given. This

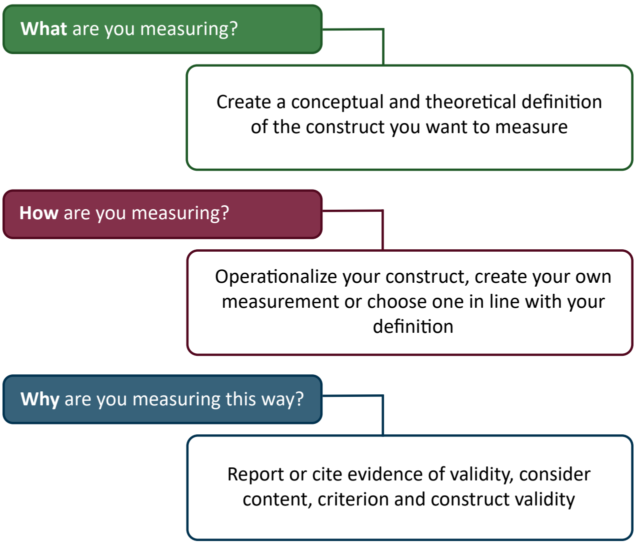

What are you measuring?
Create a conceptual and theoretical definition of the construct you want to measure
How are you measuring?
Operationalize your construct, create your own measurement or choose one in line with your definition
Why are you measuring this way?
Report or cite evidence of validity, consider content, criterion and construct validity

Fig. 2. Model of a measurement selection process

includes considerations of content, criterion and construct validity [18]. When using a validated questionnaire, these can be found in validation studies of those questionnaires. However, self-developed questionnaires must be validated according to best practice guidelines within the study in which its results are reported and interpreted.

6.3.4 Summary. The answers to all questions posed in our model of the measurement selection process should be reported in each study measuring at least one construct using a survey scale. Ideally, this should result in the transparent reporting of information in regards to definition, theory, administration, operationalization, validation and justification of a measurement used.

As a result of reporting all of these considerations, researchers can achieve transparent measurement reporting. Employing a prescriptive model of measurement selection should improve the current state of transparency, especially for presently neglected aspects such as definitions of constructs and justifications for selected measurements.

6.4 Limitations & Future Research

We note a few limitations associated with this study. Firstly, for our systematic literature review we solely analyzed the literature published in one year of the ACM CHI PLAY conference. Namely, the latest one of CHI PLAY 2020. However, we believe for our purposes of providing a current state of transparency in measurement reporting, it surmises an appropriate sample. Further, we chose CHI PLAY 2020 as we expect research to improve over time and we expect the most recent year to be the most transparent year in measurement reporting. This should provide a fair basis for issues in regards to the current state of transparency in measurement reporting

As with all literature reviews, the possibility exists for coder error or individual differences in how specific instances are understood and coded, even with especially diligent coding. As this study does not attempt to single out individual papers or make examples of them, but rather show an overall trend, any error which might have occurred should be within margin. However, in order to mitigate these concerns we assessed inter-rater reliability.

Further, this review was not preregistered, however the code book followed Flake and Fried's methodology closely and all changes to this have been detailed in this paper under the code book section. The code book and coding are further in full available on https://osf.io/4xz2v/.

For our purposes we did not differentiate between atheoretical and theoretical measures as defined by previous research [18]. Although it can be criticized that we judged measures which are supposedly atheoretical on whether they are rooted in theory, we found no paper explicitly state that certain measures are atheoretical. As such it would be our responsibility of judging the intent of authors. This was something we expressly attempted to avoid with our focus on transparency and therefore this consideration is out of scope for this study.

In regards to future research we believe the continued publication of meta-research such as systematic literature reviews, especially with a focus on popular methodologies will aid in both the quality and the ensured progress of the research field.

We further believe it is imperative for PX research to build discussions surrounding the conceptual and theoretical underpinnings of central constructs, with the eventual end goal of a conceptual consensus. Previous research has already noted the need for PX research to move beyond the constructs and methods we can 'borrow' from psychology and other UX to give us legitimacy [52]. Instead, we need to establish new language to discuss games more holistically. For this, researchers need to carefully consider which constructs, theories and measurements they borrow from psychology and other related fields. Further, we find it necessary for PX to grow its own vocabulary and theoretical underpinnings in addition borrowing conceptual and theoretical understandings from established and previously used survey-scales. To achieve this goal, explicit discussions of constructs and theory are a necessary future step.

Lastly, investigations into the validity of questionnaires as well as the creation of new, well validated questionnaires on the basis of theory, will continue to be important to ensure the high quality of PX research published at CHI PLAY.

7 CONCLUSION

The measurement of theoretical concepts, or constructs, is an integral part of PX research. The use of survey scales is a common way for researchers to gain data about these constructs through measuring them. Transparency in regards to how these measurements were carried out is vital to judge the validity of a study. However, in our methodological review of 24 full papers we found that in many instances of measurements, information in regards to definition of a construct and justification as to why a certain measurement was selected, was lacking. In comparison, details in regards to operationalization of a construct were commonly included in a paper. To address these issues we present findings as to how they can threaten the validity of a study and introduce a prescriptive model of the measurement selection process. Following the model should encourage and aid researchers in treating the measurement of constructs with the weight it deserves and report on their methods transparently. This will not only improve the quality of single studies, but rather aid in generalizing findings and future meta-analytical work in PX research.

DATA AVAILABILITY

The data for this systematic review is openly available under https://osf.io/4xz2v/. Including the full coding of all constructs, the coding of the cited measures and their validation, the citations for all coded papers, the citations for the cited measurements, the code book, a list of all the constructs found and R scripts used for analysis.

AUTHOR CONTRIBUTION

LFA, FB and SACP conceived the initial idea. LFA designed the literature review and the initial code book. SACP and FB gave feedback to the code book's iterations. LFA tested the initial and iterated code books and created iterations until the final code book was finished. LFA, SACP and LW coded the CHI PLAY 2020 papers. LFA wrote the draft. LFA, SACP, LW, KO and FB contributed to the final version.

ACKNOWLEDGEMENTS

We thank the reviewers for their helpful feedback. Special thanks to Laetitia Woelfle and Nestor for their support during the writing of this paper. This work was funded internally.

REFERENCES

[1] Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (2020), 102–370. https://doi.org/10.1016/j.ijhcs.2019.102370

[2] CHI PLAY ACM. 2021. Contribution Types and Evaluation Criteria. https://chiplay.acm.org/2021/table-1/

[3] John Antonakis, Nicolas Bastardoz, Philippe Jacquart, and Boas Shamir. 2016. Charisma: An Ill-Defined and Ill-Measured Gift. *Annual Review of Organizational Psychology and Organizational Behavior* 3, 1 (Mar. 2016), 293–319. https://doi.org/10.1146/annurev-orgpsych-041015-062305

[4] Mark Appelbaum, Harris Cooper, Rex B. Kline, Evan Mayo-Wilson, Arthur M. Nezu, and Stephen M. Rao. 2018. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist* 73, 1 (2018), 3–25. https://doi.org/10.1037/amp0000191

[5] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (US). 1999. *Standards for educational and psychological testing*. Amer Educational Research Assn, Washington, DC, USA.

[6] Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2689–2698. https://doi.org/10.1145/1978942.1979336

[7] Adam E. Barry, Beth Chaney, Anna K. Piazza-Gardner, and Enmanuel A. Chavarria. 2014. Validity and Reliability Reporting Practices in the Field of Health Education and Behavior: A Review of Seven Journals. *Health Education & Behavior* 41, 1 (2014), 12–18. https://doi.org/10.1177/1090198113483139

[8] Denny Borsboom. 2006. The attack of the psychometricians. *Psychometrika* 71, 3 (Sep. 2006), 425–440. https://doi.org/10.1007/s11336-006-1447-6

[9] Katreen Boustani, Anne C. Tally, Yu Ra Kim, and Christena Nippert-Eng. 2020. Gaming the Name: Player Strategies for Adapting to Name Constraints in Online Videogames. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 120–131. https://doi.org/10.1145/3410404.3414259

[10] Florian Brühlmann and Elisa D. Mekler. 2018. Surveys in Games User Research. In *Games User Research*, Anders Drachen, Pejman Mirza-Babaei, and Lennart Nacke (Eds.). Oxford University Press, Oxford, Chapter 9, 141–162.

[11] Florian Brühlmann and Gian-Marco Schmid. 2015. How to Measure the Game Experience?: Analysis of the Factor Structure of Two Questionnaires. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1181–1186. https://doi.org/10.1145/2702613.2732831

[12] Marcus Carter, John Downs, Bjorn Nansen, Mitchell Harrop, and Martin Gibbs. 2014. Paradigms of Games Research in HCI: A Review of 10 Years of Research at CHI. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play* (Toronto, Ontario, Canada) *(CHI PLAY '14)*. Association for Computing Machinery, New York, NY, USA, 27–36. https://doi.org/10.1145/2658537.2658708

[13] David Chan. 2009. So why ask me? Are self-report data really that bad. In *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, Charles E Lance and Robert J Vandenberg (Eds.). Taylor & Francis, Abingdon, UK, Chapter 13, 309–336.

[14] Lewis L. Chuang and Ulrike Pfeil. 2018. *Transparency and Openness Promotion Guidelines for HCI*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3170427.3185377

[15] James M. Conway and Charles E. Lance. 2010. What Reviewers Should Expect from Authors Regarding Common Method Bias in Organizational Research. *Journal of Business and Psychology* 25, 3 (May 2010), 325–334. https://doi.org/10.1007/s10869-010-9181-6

[16] John Dawes. 2008. Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point and 10-Point Scales. *International Journal of Market Research* 50, 1 (Jan. 2008), 61–104. https://doi.org/10.1177/147078530805000106

[17] Edward L. Deci and Richard M. Ryan. 2003. Intrinsic motivation inventory. *Self-determination theory* 267 (2003).

[18] Robert F. DeVellis. 2016. *Scale development: Theory and applications*. Vol. 26. Sage publications, Los Angeles, CA, USA.

[19] Florian Echtler and Maximilian Häußler. 2018. *Open Source, Open Science, and the Replication Crisis in HCI*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3188395

[20] Matthias Egger, George Davey Smith, and Jonathan AC Sterne. 2001. Uses and abuses of meta-analysis. *Clinical Medicine* 1, 6 (2001), 478. https://doi.org/10.7861/clinmedicine.1-6-478

[21] Panteleimon Ekkekakis and James A. Russell. 2013. *The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511820724

[22] Jessica Kay Flake and Eiko I. Fried. 2020. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science* 3, 4 (2020), 456–465. https://doi.org/10.1177/2515245920952393

[23] Saul Greenberg and Bill Buxton. 2008. Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 111–120. https://doi.org/10.1145/1357054.1357074

[24] Olivia Guest and Andrea E. Martin. 2021. How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science* 16, 4 (2021), 789–802. https://doi.org/10.1177/1745691620970585 arXiv:https://doi.org/10.1177/1745691620970585

[25] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The Extent and Consequences of P-Hacking in Science. *PLOS Biology* 13, 3 (Mar. 2015), 1–15. https://doi.org/10.1371/journal.pbio.1002106

[26] Kenneth D Hopkins. 1998. *Educational and psychological measurement and evaluation*. Pearson, London, UK.

[27] Kasper Hornbæk. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies* 64, 2 (2006), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

[28] Wijnand IJsselsteijn, Yvonne De Kort, and Karolien Poels. 2013. The Game Experience Questionnaire. Eindhoven: Technische Universiteit Eindhoven.

[29] Daniel Johnson, M. John Gardner, and Ryan Perry. 2018. Validation of two game experience scales: The Player Experience of Need Satisfaction (PENS) and Game Experience Questionnaire (GEQ). *International Journal of Human-Computer Studies* 118 (2018), 38 – 46. https://doi.org/10.1016/j.ijhcs.2018.05.003

[30] E. F. Juniper. 2009. Validated questionnaires should not be modified. *European Respiratory Journal* 34, 5 (Oct. 2009), 1015–1017. https://doi.org/10.1183/09031936.00110209

[31] Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. 2018. Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 257–270. https://doi.org/10.1145/3242671.3242683

[32] Effie L.-C. Law, Paul van Schaik, and Virpi Roto. 2014. Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies* 72, 6 (2014), 526 – 541. https://doi.org/10.1016/j.ijhcs.2013.09.006 Interplay between User Experience Evaluation and System Development.

[33] James R. Lewis. 2014. Usability: Lessons Learned … and Yet to Be Learned. *International Journal of Human-Computer Interaction* 30, 9 (2014), 663–684. https://doi.org/10.1080/10447318.2014.930311

[34] Elizabeth F. Loftus and Guido Zanni. 1975. Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society* 5, 1 (Jan. 1975), 86–88. https://doi.org/10.3758/bf03336715

[35] Scott B. MacKenzie. 2003. The Dangers of Poor Construct Conceptualization. *Journal of the Academy of Marketing Science* 31, 3 (Jun. 2003), 323–326. https://doi.org/10.1177/0092070303031003011

[36] Regan Mandryk. 2020. PACM Statement. https://chiplay.acm.org/2021/pacm/

[37] Andrew Maul. 2017. Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives* 15, 2 (2017), 51–69. https://doi.org/10.1080/15366367.2017.1348108

[38] Kent L. Norman. 2013. GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers* 25, 4 (Mar. 2013), 278–283. https://doi.org/10.1093/iwc/iwt009

[39] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A.

McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews* 10, 1 (29 Mar 2021), 89. https://doi.org/10.1186/s13643-021-01626-4

[40] Elazar J Pedhazur and Liora Pedhazur Schmelkin. 2013. *Measurement, design, and analysis: An integrated approach.* psychology press, Hove, East Sussex, UK.

[41] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. *A Bermuda Triangle? A Review of Method Application and Triangulation in User Experience Evaluation.* Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3173574.3174035

[42] Karolien Poels, Yvonne De Kort, and Wijnand IJsselsteijn. 2007. D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games. Eindhoven: Technische Universiteit Eindhoven.

[43] Shaghayegh Roohi, Asko Relas, Jari Takatalo, Henri Heiskanen, and Perttu Hämäläinen. 2020. Predicting Game Difficulty and Churn Without Players. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20).* Association for Computing Machinery, New York, NY, USA, 585–593. https://doi.org/10.1145/3410404.3414235

[44] R Nevitt Sanford, Theodor W Adorno, Else Frenkel-Brunswik, and Daniel J Levinson. 1950. The measurement of implicit antidemocratic trends. *The authoritarian personality* (1950), 222–279.

[45] Liam P. Satchell, Dean Fido, Craig A. Harper, Heather Shaw, Brittany Davidson, David A. Ellis, Claire M. Hart, Rahul Jalil, Alice Jones Bartoli, Linda K. Kaye, Gary L. J. Lancaster, and Melissa Pavetich. 2020. Development of an Offline-Friend Addiction Questionnaire (O-FAQ): Are most people really social addicts? *Behavior Research Methods* 52, 5 (24 Sep 2020). https://doi.org/10.3758/s13428-020-01462-9

[46] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (2005), 695–729. https://doi.org/10.1177/0539018405058216

[47] Donald Sharpe. 2013. Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods* 18, 4 (2013), 572–582. https://doi.org/10.1037/a0034177

[48] Pamela Shoemaker, James W. Tankard, and Dominic L. Lasorsa. 2004. Theoretical concepts: The building blocks of theory. In *How to Build Social Science Theories.* SAGE Publications, Inc., Los Angeles, CA, 15–36. https://doi.org/10.4135/9781412990110

[49] Brent D. Slife, Casey D. Wright, and Stephen C. Yanchar. 2016. Using Operational Definitions in Research: A Best-Practices Approach. *The Journal of Mind and Behavior* 37, 2 (2016), 119–139. http://www.jstor.org/stable/44631540

[50] Velvet Spors, Gisela Reyes Cruz, H. R. Cameron, Martin Flintham, Pat Brundell, and David Murphy. 2020. Plastic Buttons, Complex People: An Ethnomethodology-Informed Ethnography of a Video Game Museum. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20).* Association for Computing Machinery, New York, NY, USA, 594–605. https://doi.org/10.1145/3410404.3414234

[51] April Tyack and Elisa D. Mekler. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3313831.3376723

[52] April Tyack and Elisa D. Mekler. 2021. Off-Peak: An Examination of Ordinary Player Experience. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 115, 12 pages. https://doi.org/10.1145/3411764.3445230

[53] Elina Vessonen. 2020. Respectful operationalism. *Theory & Psychology* 31, 1 (2020), 84–105. https://doi.org/10.1177/0959354320945036

[54] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. 2020. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play.* ACM, New York, NY, USA, 4–18. https://doi.org/10.1145/3410404.3414229

[55] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376448

[56] Larry J. Williams and Stella E. Anderson. 1994. An Alternative Approach to Method Effects by Using Latent-Variable Models: Applications in Organizational Behavior Research. *Journal of Applied Psychology* 79, 3 (1994), 323 – 331. https://doi.org/10.1037/0021-9010.79.3.323

[57] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (Apr. 2016), 38–44. https://doi.org/10.1145/2907069

Received February 2021; revised June 2021; accepted July 2021

Independent Validation of the Player Experience Inventory: Findings from a Large Set of Video Game Players

SEBASTIAN A. C. PERRIG*, NICOLAS SCHAROWSKI, FLORIAN BRÜHLMANN, NICK VON FELTEN, KLAUS OPWIS, and LENA FANYA AESCHBACH, Center for General Psychology and Methodology, University of Basel, Switzerland

Measuring the subjective experience of digital game players is essential to player experience research. Recently, the Player Experience Inventory (PXI) was developed, which assesses both functional and psychosocial consequences of digital gameplay. We present a pre-registered independent online study with a large sample to provide additional evidence of psychometric quality for the PXI. Responses from 1518 participants were collected, rating a recent or memorable experience playing a digital game using the PXI and related measures. While our results from standard psychometric reliability and validity analyses generally favored the PXI, we also identified challenges with the immersion construct. Further, we find a ten-factor model, or alternatively, an 11-factor should enjoyment be measured, to fit our collected data best. In sum, the PXI is a valuable tool to measure a variety of constructs central to player experience.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; **HCI theory, concepts and models**; • **Applied computing** → **Computer games**.

Additional Key Words and Phrases: Player experience, Player Experience Inventory, Games user research, Questionnaires, Survey scales, Scale validation, Measurement instrument, Psychometrics

ACM Reference Format:
Sebastian A. C. Perrig, Nicolas Scharowski, Florian Brühlmann, Nick von Felten, Klaus Opwis, and Lena Fanya Aeschbach. 2024. Independent Validation of the Player Experience Inventory: Findings from a Large Set of Video Game Players. In *CHI 2024: ACM CHI Conference on Human Factors in Computing Systems, May 11–16, 2024, Honolulu, HI*. ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In Games User Research (GUR), employing self-reports to measure players' subjective experiences is a popular method [10]. However, few properly validated survey scales exist for measuring player experience (PX), especially with a strong focus on providing actionable insights for practical game design. To fill this gap, Abeele et al. [1] developed the Player Experience Inventory (PXI), a 30-item survey scale based on means-end theory [25, 54]. Since its initial development, the PXI has been translated into German [24], and both a short version [27] and a benchmark [26] for the scale have been created. Beyond this work on the PXI by people associated with its original authors, no independent evaluation of the PXI has occurred [65]. Furthermore, the samples used to develop and validate the PXI were below commonly recommended sample sizes for essential scale evaluation and validation methods, such as confirmatory factor analysis (CFA) [40]. In addition, past validation work on the PXI has primarily sampled students or volunteers. However, other samples, such as crowd-sourced participants, have seen increased use in research [e.g., 11]. Similarly, participants were predominately young men, something acknowledged as a limitation by the original authors, and thus not representative

[*] Corresponding Author.

of general player demographics, where an equal share of men and women report playing digital games [20]. Thus, it remains to be seen whether the PXI also applies to more diverse populations.

Given that the quality of research findings depends on the reliability and validity of the methods used [3], it is of utmost importance for researchers studying digital games to have an adequately validated scale for measuring players' experiences. The PXI is a promising scale, with its focus on both the psychosocial consequences of gameplay and the functional consequences of game mechanics. However, an independent investigation into the quality of the PXI has yet to be conducted because previous studies on the PXI were always run by people directly associated with the original team of researchers behind the PXI. Motivated by this research gap, we assessed the psychometric quality of the PXI in a pre-registered online survey following current best practices for scale quality investigation. Thus, this work's contribution is a large-sample independent validation of the PXI. Data from 1518 crowd-sourced participants was collected, who were asked to rate a recent or memorable experience playing a digital game using the PXI and related scales. Results generally demonstrated good psychometric quality for the PXI and supported the proposed ten-factor theoretical model behind the scale. Standard reliability and validity measures mainly favored the PXI but indicated room for improvement regarding certain constructs. In particular, the construct of immersion was negatively salient in several respects. Overall, the findings of this study demonstrated that the PXI is a reliable and valid tool for measuring players' experience with digital games, contributing to a more accurate measurement of the gaming experience.

2 RELATED WORK

2.1 The PXI

The PXI is a 30-item survey scale developed and validated based on input from 64 GUR experts in two iterations and data collected from 529 players across five studies. The PXI was designed to measure digital games' psychosocial and functional consequences. Functional consequences are defined as "the immediate and tangible consequences that are experienced directly by consumers, during the use of the product" [1, p. 3]. Psychosocial consequences, in contrast, "exceed the immediate usage level and reach into the social or, psychological level" [1, p. 3-4]. While there are numerous scales to gauge psychosocial consequences in PX research, the measurement of functional consequences is unique to the PXI [1].

The PXI measures ten different constructs, five each for the functional and psychosocial consequences of playing digital games. The constructs and their respective definitions are presented in Table 1. Per construct, three items are used, to which participants' answers are recorded on a seven-point Likert-type response scale, ranging from -3 (Strongly disagree) to +3 (Strongly agree). For functional consequences, the scale measures the following constructs: *ease of control, challenge, progress feedback, goals and rules*, and *audiovisual appeal*. The specific constructs of the psychosocial consequences measured with the PXI are *meaning, immersion, mastery, curiosity*, and *autonomy*. Beyond the ten constructs of the PXI, there is an additional construct, namely *enjoyment*, suggested by the authors to be measured alongside the PXI with three dedicated items but not considered part of the actual scale. In the context of media entertainment, enjoyment has been described as "an individual's positive response towards media technology and its content" [[68] in 48, p. 927].

Graf et al. [24] translated the PXI into German and validated the translated version in an online study (N = 506). Results showed that the translated version had good psychometric properties, although there was room for improvement concerning the scale's discriminant validity. Besides the original PXI, a short version exists, which consists of 11 items. With one item per construct, including enjoyment, the miniPXI [27] was developed across three studies based on data and insights from 15 experts and 628 digital game players. In addition to these scale versions, Haider et al. [26] developed the PXI bench, an online tool for the analysis and comparison of PXI response data, which can be accessed on the PXI's official website.[1] However, no additional investigation into the psychometric quality of the PXI has occurred since the original validation. Thus, there is a need for additional validation of the PXI to see if the initial evidence for the scale's quality can be reproduced.

Beyond the general psychometric quality, the theoretical model behind the PXI also calls for further inquiry. In their paper, the PXI's authors employed factor analyses to test a ten-factor model for the PXI, with a dedicated factor per construct [1]. In this regard, it was left open what the psychometric quality of the enjoyment items was and how an enjoyment factor might fit into this model, if at all. Furthermore, the PXI's authors performed a mediation analysis to investigate the theoretical model and, thus, the relationship between the functional and psychosocial consequences, as well as game enjoyment. However, they did not report detailed results on how a model considering the ten constructs, the consequences, and game enjoyment would perform. Given the scale's name, one might further expect the PXI to measure an overall factor of PX, although the authors have never suggested this. Nevertheless, all of this leaves the question of whether a simple ten-factor model best fits the scale or if alternative models, including higher-order factors, such as for the consequences or overall PX, are better suited.

2.2 Usage of the PXI

As part of our initial investigation into the PXI, we first conducted a literature review of the peer-reviewed articles that cited the original papers on the English and German PXI [i.e., 1, 24, 66] as of March 2023 ($N = 45$). Details on the literature review can be found in the supplementary materials on OSF. We aimed to understand how the PXI has been used within academia since its publication. This aided us in gaining a bottom-up perspective of the measurement models to validate for the PXI. We found that quality investigations of the scale, even when used in a novel context or with selected dimensions, were infrequent. Further, the quantification of the scale was sometimes unclear, meaning it was uncertain how researchers averaged the response items for further calculations. However, in some instances where quantification could be assessed, we found the authors computed a general PX score. In other words, researchers would

[1] https://playerexperienceinventory.org/, last accessed on August 24, 2023.

Table 1. Constructs of the PXI, as defined in Abeele et al. [1, p. 5].

Construct	**Definition**
Functional consequences	
ease of control	"The extent to which a player finds the actions to control the game clear and intuitive"
challenge	"The extent to which the specific challenges in the game match the players skill level"
progress feedback	"The extent to which it is clear to the player how well he or she is doing in the game"
goals and rules	"The extent to which the overall objective and rules are clear to the player"
audiovisual appeal	"The extent to which a player appreciates the audiovisual styling of the game"
Psychosocial consequences	
meaning	"A sense of connecting with the game, resonating with what is important"
immersion	"A sense of immersion and cognitive absorption, experienced by the player"
mastery	"A sense of competence and mastery derived from playing the game"
curiosity	"A sense of interest and curiosity roused by the game"
autonomy	"A sense of freedom and autonomy to play the game as desired"

average all the item scores from the ten, or sometimes eleven, dimensions of the PXI into a single overall score per participant. Indeed, some researchers also described using the PXI to measure PX generally without referring to the actual factors of the scale. As such, it is important to investigate a model of the PXI that includes a general PX factor to understand whether such usage is psychometrically justified. Additionally, we found many researchers employing the suggested enjoyment items, which, as described above, have not been validated alongside the other ten factors of the scale. Another frequently measured dimension was immersion. This dimension was often used in conjunction with other scales for specific contexts, specifically virtual reality applications.

We also found that researchers administered the scale differently from the originally proposed version despite the PXI's authors stressing the importance of using the original scale and response options on their website. Instead of the -3 to +3 range of the Likert-type response scale, often a range from 1 to 7 or even 1 to 5 was employed.

2.3 The importance of independent validation

As the authors of the PXI themselves emphasized, "scale development and validation is an ongoing process" [1, p. 10]. Regarding the quality of a survey scale, researchers are typically interested in three criteria: objectivity, reliability, and validity [18]. Objectivity signifies that "any statement of fact made by one scientist should be independently verifiable by other scientists" [51, p. 6]. Reliability considers "how accurately a test measures the thing which it does measure" [39, p. 14]. Finally, validity is concerned with "whether a test really measures what it purports to measure" [39, p. 14]. Only if all three quality criteria are met can researchers have confidence in the data gathered using survey scales and, consequently, in the conclusions derived from them. While the original work on the PXI provided extensive results that speak to the quality of the scale, these results are limited to the sample and setting of the original paper. Ideally, a scale's psychometric quality should be assessed whenever it is used [22]. Because this is not always realistic, independent validation in other settings can provide valuable additional insight into the quality of a scale. Furthermore, the PXI was developed and validated based on data from predominantly young men, contrasting general demographics of digital game players [20]. This limitation was also acknowledged by its authors, who called for further studies "to assess how the PXI performs across different game audiences" [1, p. 10-11]. Thus, additional studies are needed to determine how the PXI performs in other populations, as the psychometric properties of a scale can vary considerably between different groups of people [22]. In addition, concerning the popular recruitment approach of crowd-sourcing [11], evaluation of the PXI has yet to be conducted, given that past samples mainly consisted of students and volunteers. Several previous studies [e.g., 36, 42, 49, 67] have shown that other scales proposed to measure players' experiences do not hold up under second inspection or at least require certain modifications to achieve satisfactory psychometric quality. For the PXI, such independent validation is still pending [65]. In summary, although the initial results on the psychometric quality of the PXI are promising, additional evidence is needed in order for researchers, both in industry and academia, to use the scale with confidence.

3 METHODS

A pre-registered online study was conducted to evaluate the psychometric quality of the PXI. During the online study, participants were asked to think about a digital game they recently played or know well before responding to several standardized survey scales, including the PXI. The study was reviewed and approved by the ethics committee of the authors' university and pre-registered on OSF (https://osf.io/buq5t/?view_only=5d69aed003e94ac5b04820c33fdb101a).

3.1 Measures

After choosing a game to think about, participants responded to all PXI items, in addition to several additional scales related to the PXI, namely the Player Experience of Need Satisfaction scale [PENS, 59], the AttrakDiff [30], and the interest/enjoyment subscale from the Intrinsic Motivation Inventory [IMI, 56, 58]. The selection of the additional scales was mostly based on the original work on the PXI. All items were presented in a randomized order with one individual page per scale unless stated otherwise. The exact wording of all items is provided in the supplementary materials, except for the PENS [59] due to copyright reasons. Reliability for all scales was investigated using the internal consistency coefficients α [13] and ω [46], which delivered satisfactory results ($\geq$ .70) for all scales, except for the PXI's immersion construct and the AttrakDiff's pragmatic quality [30], which fell just below the desired threshold (see subsection 4.3 for results on the PXI, and the supplementary materials on OSF for the other scales).

3.1.1 PXI. All 30 items of the English PXI were used alongside the three suggested items for enjoyment. Items were distributed across three pages, likewise to the survey on the PXI website, and responses were collected using the recommended seven-point Likert-type response scale ranging from -3 ("Strongly disagree") to +3 ("Strongly agree").

3.1.2 PENS. Participants responded to all 21 items of the PENS [59]. We chose the PENS because it was already used in the original validation for the PXI and contains several constructs related to those of the PXI: *autonomy, competence, relatedness*, and *intuitive controls*, with three items each and the construct *presence* with nine items. In the context of the PENS and self-determination theory, autonomy "concerns a sense of volition or willingness when doing a task" [[16, 17] in 59, p. 349]. Competence refers to "a need for challenge and feelings of effectance" [[15, 70] in 59, p. 349] while "[r]elatedness is experienced when a person feels connected with others" [[41, 57] in 59, p. 350]. The intuitive controls construct considers "whether [the game controls] make sense, are easily mastered, and do not interfere with once sense of being in the game" [sic, 59, p. 350]. Finally, presence describes "the sense that one is *within* the game world, as opposed to experiencing oneself as a person outside the game, manipulating controls or characters" [59, p. 350]. Responses to the PENS were collected on a seven-point Likert-type response scale from 1 ("Do not agree") to 7 ("Strongly agree").

3.1.3 AttrakDiff. As in the original PXI paper, participants responded to the AttrakDiff semantic differential scale [30]. We used the most recent 28-item version of the scale, available on the official website,[2] which measures four constructs with seven items each: *pragmatic quality* (PQ), *hedonic quality - identification* (HQ-I), *hedonic quality - stimulation* (HQ-S), and *attractiveness* (ATT). Pragmatic quality concerns attributes of a system "connected to the users' need to achieve behavioral goals" while "hedonic attributes are primarily related to the users' self" [30, p. 322]. Stimulation, alongside novelty and challenge, is considered "a prerequisite of personal development [...] which in turn is a basic human need" [30, p. 322]. Identification, on the other hand, "addresses the human need to express one's self through objects" [30, p. 322]. Finally, attractiveness "is a global assessment based on the perceived [product] qualities" [31, p. 3, translated from German]. Responses were collected on a seven-point semantic differential response scale, with, for example, the words "ugly" and "attractive" at two opposing poles.

3.1.4 IMI - interest/enjoyment. In addition, we had participants fill out the subscale for interest/enjoyment from the IMI [56, 58]. Responses to the seven items were collected on the seven-point Likert-type response scale from 1 ("Not at all true") to 7 ("Very true") recommended by the authors, and items were slightly adapted to fit the gaming context, which

[2] https://www.attrakdiff.de/, last accessed on August 24, 2023.

is commonly done [56]. We chose the IMI because it is the most frequently used scale to measure game enjoyment [48] and because enjoyment was also measured in the original PXI paper.

3.2 Procedure

Participants provided informed consent on the first page of the online survey. Next, they were given instructions for the task to be completed. Following the original PXI paper, we asked participants to recall an experience with a game they recently played or know well. For this, we used the critical incident technique, commonly used in HCI research [e.g., 6, 62], asking participants to describe the game in at least 50 words. The exact wording of the critical incident question was as follows:

"Please describe the digital game you recently played or that you remember well. Try to describe this particular game as accurately and detailed as you remember in at least 50 words, and try to be as concrete as possible. You can use as many sentences as you like."

Participants were further instructed to provide the name of the chosen game, which was then used in subsequent questions to personalize the survey. This ensured that participants would think about the described game while filling out the survey (e.g., "Please fill out the following questions for the digital game you recently played or that you remember well ([name of game])."). After the critical incident question, participants responded to several closed questions concerning the chosen game, which we adopted from the survey on the PXI website (e.g., controls used, the platform played on) before filling out the PXI and the three items for game enjoyment [1]. On the following survey pages, participants filled out the other scales in a randomized order. Next, participants provided demographic information (age, gender, country of residence, game experience, playtime). Lastly, participants could give feedback before receiving their compensation. To ensure sufficient response quality, the survey included two instructed response items [14] embedded among the survey scales and a single item for self-reported data quality [47] at the end of the survey. Completing the survey took participants an average of 11.71 minutes (SD = 5.54, min = 3.48, max = 52.92).

3.3 Pre-study

Before pre-registering the study, we tested the procedure and task with a small-sample pre-study (N = 50) to examine if participants could complete the task and if there were any major issues with the study procedure. The recruitment criteria for the pre-study were the same as for the main study (see below). Participants encountered no issues, and all responses, including the critical incident question, were satisfactory. Thus, no changes in the study procedure or the recruitment criteria were necessary. For this reason, the data from the pre-study was combined with the main sample for the analysis.

3.4 Participants

Prolific, a crowd-sourcing platform recently shown to have high data quality [19, 52], was used for recruitment. A total of 1501 participants from the United Kingdom (UK) were recruited and reimbursed £1.50 for completing the study. Participants were screened on Prolific on whether they play digital games at least occasionally. A target sample size of at least 1050 responses after data cleaning was set based on rules of thumb for structural equation modeling, recommending at least ten observations per estimated model parameter [40]. We followed recommendations by Brühlmann et al. [11] for data cleaning, filtering out participants using two instructed response items [14], a seriousness check [47], and responses to open answers. Responses from nine participants were removed based on the seriousness check, and another three responses were removed due to an incomplete or interrupted survey. Five additional participants were removed for indicating their current country of residence outside the UK, and 16 were removed based on low-quality critical incident game descriptions (e.g., repeating words to reach the word minimum, not describing a digital game, indicating that they could not accurately remember the game). The final sample, including participants from the pre-study, consisted of 1518 responses. Thus, no additional recruitment was needed to achieve the target sample size. Of the participants, 639 were women, 864 were men, nine were non-binary people, one person preferred to self-describe, and five people chose not to provide information on their gender. The average age of participants was 37.47 years ($SD = 12.18, min = 18, max = 79$). The most frequently used game platform was consoles (574 participants), followed by PC (497) and smartphones (435). Participants most frequently stated that they played the rated game alone (1086), followed by playing online with other players (481) and playing locally with others (131). Most participants used controllers to play the game (620), followed by touch controls (507) and keyboards (438).[3] The most frequently rated digital game was FIFA (mentioned 60 times), followed by Candy Crush (57), The Sims (37), Mario Kart (33), Call of Duty (30), Grand Theft Auto (30), Fortnite (27), and Minecraft (27). The most popular genres, self-reported by the participants, were puzzle games (282), action-adventure (270), and action role-playing (155). On average, players rated their game expertise at 4.84 on a seven-point response scale ($SD = 1.38, min = 1.00, max = 7.00$), and most often indicated a playtime of 5 to 10 hours per week (421 participants), 2 to 5 hours (396), and 10 to 20 hours (291).

4 RESULTS

The following section describes different forms of psychometric quality investigation for the PXI. The complete analysis can be found in the supplementary materials on OSF. The analyses mostly followed those methods used in the original work on the PXI. All results were obtained using the statistical software R [53, version 4.3.0]. Overall descriptive statistics for the collected data are presented in Table 2.

4.1 Item analysis

We began the psychometric investigation into the PXI using item analysis. We considered descriptive statistics, item difficulty and variance, discriminatory power (i.e., item-total correlation), and inter-item correlations for all 30 PXI items and the three enjoyment items. In summary, the item analysis showed no problematic values for most PXI items. However, a few items exhibited conspicuous results. Namely, descriptive statistics deviated from the other items for item immersion_1, which exhibited a lower mean and different distribution of responses compared to other items. Further, the item variances were below 1 for multiple items (see supplementary materials for details). We thus continued with the analysis while keeping those items in mind for the interpretation of further results.

4.2 Confirmatory factor analyses

As pre-registered, we next performed multiple CFAs to investigate the model fit of the PXI. We tested multiple models for the following reasons. First, we encountered different conceptualizations regarding the application of the PXI in our literature review on the current usage of the PXI. Second, the original work on the scale likewise offers multiple conceptualizations, specifically concerning the distinction of the PXI's constructs into functional and psychosocial consequences. Given that these applied or proposed models differed regarding the inclusion of certain higher-order factors for the functional and psychosocial consequences and/or for PX, we conducted multiple CFAs corresponding to these conceptualizations, as recommended by Brown [9]. The multivariate normality assumption was not met, tested

[3] More than one selection was possible for the game descriptions; hence, the total adds up to more than 1518.

Table 2. Descriptive statistics for all collected measures.

Construct	Mean	SD	Min	Max
PXI meaning	1.39	1.15	-3.00	3.00
PXI curiosity	1.71	1.17	-3.00	3.00
PXI mastery	1.74	0.91	-3.00	3.00
PXI autonomy	1.68	1.17	-3.00	3.00
PXI immersion	1.50	1.04	-3.00	3.00
PXI progress feedback	1.90	0.97	-3.00	3.00
PXI audiovisual appeal	2.20	0.85	-3.00	3.00
PXI challenge	1.60	0.96	-2.67	3.00
PXI ease of control	2.10	0.79	-2.00	3.00
PXI goals and rules	2.40	0.67	-1.33	3.00
PXI enjoyment	2.41	0.70	-2.67	3.00
AttrakDiff HQ-S	1.26	1.03	-2.86	3.00
AttrakDiff HQ-I	1.08	0.95	-3.00	3.00
AttrakDiff ATT	1.86	0.88	-3.00	3.00
AttrakDiff PQ	1.07	0.82	-2.29	3.00
PENS autonomy	5.26	1.28	1.00	7.00
PENS competence	5.61	0.99	1.00	7.00
PENS relatedness	3.64	1.61	1.00	7.00
PENS presence	3.96	1.45	1.00	7.00
PENS intuitive controls	5.71	1.04	1.00	7.00
IMI interest/enjoyment	5.95	0.91	1.86	7.00

Note: Responses could range from -3 to +3 for the PXI and AttrakDiff, and from 1 to 7 for the other scales.

using the Henze-Zirkler test [32] and Mardia's test [44]. We thus chose to use a robust maximum likelihood estimator with a Yuan-Bentler scaling correction for all CFAs, which is recommended for non-normal data and reduces the risk of Type I error [9]. In all analyses, the factor loading for the first indicator of each latent variable was constrained to one, as is standard procedure when defining a metric for each factor [9, 40, 55, 63]. For the judgment of model fit, we opted for the same criteria used during the original PXI validation (see Table 3), combining multiple criteria to improve the acceptability of Type I and Type II error rates [[34] in 9].

Based on the information provided by the original authors of the PXI and the findings of our literature review on how the PXI is currently used in research, we investigated the fit of five different models to the collected data.

Table 3. Cut-off criteria for model fit indices considered, as used in Abeele et al. [1] based on [12, 33, 34, 64].

	Acceptable	Excellent
χ^2/df	< 5	< 2
CFI	> .90	> .95
TLI	> .90	> .95
RMSEA	< .08	< .06
SRMR	< .09	< .08

Table 4. Fit indices for CFA models of the PXI. Models 1 and 5 were assessed without higher-order factors, and models 2, 3 & 4 included varying higher-order factors.

Tested model	χ^2	**df**	**p-value** χ^2	**RMSEA**	**SRMR**	**CFI**	**TLI**	χ^2/df
1) 10 factors (original PXI)	1053.213	360	< .001	.041	.041	.956	.946	2.926
2) 10 factors + 2 factors consequences	1866.132	394	< .001	.057	.079	.906	.896	4.736
3) 10 factors + 1 factor PX	2000.001	395	< .001	.060	.079	.897	.886	5.063
4) 10 factors + 2 factors consequences + 1 factor PX	1861.395	393	< .001	.057	.079	.906	.896	4.736
5) 11 factors, incl. enjoyment	1278.195	440	< .001	.041	.040	.955	.946	2.905

Note: Robust values are reported wherever possible.

Following the factors proposed in the original work on the PXI, we started with a ten-factor model for the 30 PXI items, with one factor each for the PXI's subscales. All items were specified to load on their designated factor. In addition, we investigated a model with two higher-order factors, one each for the functional and psychosocial consequences, upon which the five respective factors of the PXI constructs loaded. This model was based on the originally proposed theoretical structure behind the PXI, which suggests further separating the ten factors of the scale into functional and psychosocial consequences. Two further models included an overall general factor for PX, once with and once without the higher-order factors for the functional and psychosocial consequences. We tested those models with an overall PX factor based on our findings from the literature review that some authors form an overall score using the items of the PXI. Finally, an additional 11-factor model was tested, including a factor for the three enjoyment items. Items were specified to load on their designated factor, following the theoretical structure proposed by the PXI's authors. All results from the CFAs are presented in Table 4.

Results from the CFAs indicated an acceptable to excellent fit of the models without higher-order factors to the data, both with and without the enjoyment items (three criteria excellent, two acceptable), judging by the cut-off criteria for model fit as used in the original work on the PXI (see Table 3). The model including higher-order factors for the functional and psychosocial consequences exhibited slightly worse but still mostly acceptable to excellent model fit statistics (two criteria excellent, two acceptable, one not acceptable). Regarding the model including a higher-order factor for PX in addition to the consequences, the fit was also slightly worse compared to the models without higher-order factors (two criteria excellent, two acceptable, one not acceptable), and a warning suggested that the model might not be identified. For the model including just a higher-order factor for PX, without consequences, the fit indices mostly fell just outside of the desired thresholds (one criterion excellent, one acceptable, three not acceptable).

In addition to comparing multiple fit indices to judge model fit, we used χ^2 difference tests to see if the model fit would differ significantly among the three nested models one through three. Given the warning for model four (not identified), we did not include it in this analysis. Results are reported in Table 5. In general, the results were in line with the findings thus far. Given that the χ^2 difference test was significant, the "larger" model with more freely estimated parameters (model one) fit the data better than the "smaller" models (two and three) in which the parameters in question were fixed [9, 69]. Thus, the ten-factor model without higher-order factors fit the data best (model one), followed by the model including two factors for the consequences (model two). In contrast, the third model with an overall factor for PX fit the data worst. Finally, both the Akaike information criterion [AIC, 4] and the Bayesian information criterion [BIC, 61] also reported in Table 5 favored the 10-factor model over all other models [9].

Table 5. AIC, BIC, and results from χ^2 difference tests for the comparison of the nested models one, two, and three of the PXI.

Tested model	**df**	**AIC**	**BIC**	χ^2	χ^2**-difference**	**df-difference**	**p-value** χ^2
1) 10 factors (original PXI)	360	120417	120977	1406.263			
2) 10 factors + 2 factors consequences	394	121437	121815	2494.006	805.946	34	< .001
3) 10 factors + 1 factor PX	395	121627	122000	2685.626	49.818	1	< .001
4) 10 factors + 2 factors consequences + 1 factor PX	393	121439	121823				
5) 11 factors, incl. enjoyment	440	127193	127838				

Note: The χ^2 column contains standard (non-scaled) test statistics. The χ^2 difference test used a Satorra and Bentler [60] correction.

4.3 Reliability

We calculated both coefficients α [13] and ω [46], including 95% confidence intervals, as indicators of reliability, based on recommendations by Dunn et al. [21]. Table 6 contains all values for both coefficients, separated by PXI subscale and for the overall scale. All values were above .70, indicating adequate internal consistency [23], except for the immersion subscale, which was just below the desired threshold.

4.4 Convergent and discriminant validity

In the original PXI paper, the convergent and discriminant validity of the constructs was assessed through composite reliability (CR), average variance extracted (AVE), and maximum and shared variance (MSV). We followed this procedure. Values for CR, AVE, and MSV were calculated based on the 11-factor CFA. All results are presented in Table 6. We also calculated CR, AVE, and MSV based on a ten-factor CFA without enjoyment, which yielded comparable results (see supplementary materials). Results were interpreted as follows [29]: CR should be ≥ .70 as evidence for reliability. Concerning a construct's convergent validity, the AVE should be ≥ .50. For discriminant validity, a construct's AVE

Table 6. Coefficients α and ω for the PXI, including 95% confidence intervals, as well as composite reliability (CR), average variance extracted (AVE), and maximum shared variance (MSV) for the subscales.

	Coefficient α	**Coefficient** ω	**CR**	**AVE**	**MSV**
overall	.91 [.90, .92]	.90 [.89, .91]			
overall (with enjoyment)	.92 [.92, .93]	.90 [.89, .91]			
meaning	.82 [.80, .84]	.83 [.81, .85]	.82	.62	.58
curiosity	.87 [.85, .89]	.87 [.85, .89]	.87	.69	.39
mastery	.72 [.69, .75]	.72 [.69, .75]	.74	.47	.45
autonomy	.84 [.82, .86]	.84 [.82, .86]	.84	.64	.37
immersion	.68 [.65, .71]	.68 [.65, .72]	.75	.40	.61
progress feedback	.78 [.75, .81]	.79 [.76, .82]	.78	.55	.44
audiovisual appeal	.87 [.85, .89]	.87 [.85, .89]	.87	.69	.57
challenge	.71 [.67, .74]	.71 [.67, .75]	.70	.44	.42
ease of control	.73 [.70, .76]	.74 [.70, .77]	.74	.48	.46
goals and rules	.76 [.73, .79]	.77 [.73, .79]	.77	.53	.46
enjoyment	.88 [.86, .89]	.88 [.86, .89]	.88	.70	.61

Note: Both coefficient ω and CR are reported, although these terms refer to the same statistic, given that different methods were used to calculate these values (i.e., CR based on 11-factor CFA and ω with the MBESS R package [37, 38]).

Table 7. Square root of AVE (in bold) and inter-construct correlations for discriminant validity.

	meaning	curiosity	mastery	autonomy	immersion	progress	appeal	challenge	control	goals	enjoyment
meaning	**.79**										
curiosity	.63	**.83**									
mastery	.63	.47	**.69**								
autonomy	.61	.54	.53	**.80**							
immersion	.76	.59	.58	.54	**.63**						
progress feedback	.27	.33	.51	.25	.27	**.74**					
audiovisual appeal	.59	.56	.52	.51	.64	.34	**.83**				
challenge	.40	.33	.65	.41	.41	.37	.40	**.66**			
ease of control	.23	.17	.67	.25	.24	.50	.37	.44	**.69**		
goals and rules	.26	.20	.54	.20	.35	.66	.38	.41	.68	**.73**	
enjoyment	.67	.59	.64	.58	.78	.37	.75	.53	.39	.46	**.84**

should be larger than its MSV, and the square root of the AVE of a construct should be greater than any inter-construct correlation (reported in Table 7).

Regarding CR, all subscales met the desired value of ≥ .70. AVE was good for more than half of the PXI's constructs but slightly below the desired value of ≥ .50 for mastery, immersion, challenge, and ease of control. MSV values were smaller than AVE for all constructs except immersion, indicating predominantly good discriminant validity. Further evidence for the PXI's discriminant validity was also shown by the results in Table 7, as the square root of most constructs' AVE was greater than the inter-construct correlations, although at times just barely. Only the inter-construct correlations between immersion and three other constructs, meaning, audiovisual appeal, and enjoyment, were greater than the square root of immersion's AVE.

4.5 Criterion validity

To assess the criterion validity of the PXI constructs, we considered bivariate correlations (Pearson's r) between the PXI and selected constructs of the other scales as indicators of criterion validity (see Table 8). The mapping of the PXI's constructs to those measured with the other scales was taken from the original PXI paper.[4] Furthermore, we considered the correlation between the PXI enjoyment items and the IMI. Results mainly showed strong positive correlations between the PXI constructs and their mapped counterparts, as expected based on the original PXI paper. However, the PXI construct challenge correlated only moderately with the construct of competence from the PENS, while meaning showed a moderate positive correlation with attractiveness from the AttrakDiff. At the same time, progress feedback and goals and rules showed only weak correlations with pragmatic quality from the AttrakDiff. Thus, correlation results mostly favored the PXI constructs' criterion validity.

Finally, we calculated a hybrid structural equation model to investigate the theoretical relationship between the psychosocial consequences, the functional consequences, and the enjoyment items of the PXI. Based on the original work on the PXI, we expected the following relationships:

- The functional consequences positively predict enjoyment.
- The functional consequences positively predict the psychosocial consequences.
- The psychosocial consequences positively predict enjoyment.
- The effect of the functional consequences on enjoyment is mediated via the psychosocial consequences.

[4] In the original PXI paper, an older version of the AttrakDiff was used, which has a single item for attractiveness and a single item for beauty. The version we used (i.e., the current version from the official website on the scale) has seven items for attractiveness, which we used for the correlations instead of the two single items.

Table 8. Correlations between constructs of the PXI and conceptually related constructs from the PENS, AttrakDiff, and IMI, including 95% confidence intervals.

PXI construct	Related construct	Pearson's r
meaning	AttrakDiff attractiveness	.497 [.458, .534]
curiosity	PENS presence	.542 [.506, .577]
mastery	PENS competence	.690 [.662, .715]
autonomy	PENS autonomy	.739 [.716, .761]
immersion	PENS presence	.558 [.523, .592]
progress feedback	AttrakDiff pragmatic quality	.270 [.223, .316]
audiovisual appeal	AttrakDiff attractiveness	.557 [.521, .591]
challenge	PENS competence	.447 [.405, .486]
ease of control	PENS intuitive controls	.645 [.615, .674]
goals and rules	AttrakDiff pragmatic quality	.296 [.249, .341]
enjoyment	IMI interest/enjoyment	.801 [.783, .819]

Note: All correlations were significant at $p < .001$.

The model displayed in Figure 1 exhibited an acceptable to excellent model fit [$\chi^2(482) = 2215.917$, $p < .001$, $RMSEA = .056$, $SRMR = .072$, $CFI = .908$, $TLI = .899$, $\chi^2/df = 4.597$]. All relationships were as expected and comparable to those reported in the original PXI paper. Functional consequences positively predict psychosocial consequences ($\beta = 0.807$, $SE = 0.043$, $z = 18.778$, $p < .001$, 95% CI [0.723, 0.891]) and psychosocial consequences positively predict enjoyment ($\beta = 0.513$, $SE = 0.099$, $z = 5.165$, $p < .001$, 95% CI [0.318, 0.708]). The indirect effect supported psychosocial consequences to be a mediator between functional consequences and enjoyment ($\beta = 0.414$, $SE = 0.065$, $z = 6.335$, $p < .001$, 95% CI [0.286, 0.542]). Moreover, a significant direct effect of functional consequences on enjoyment was found ($\beta = 0.416$, $SE = 0.103$, $z = 4.030$, $p = .003$, 95% CI [0.214, 0.618]). The total combined effect of functional and psychosocial consequences on enjoyment was also significant ($\beta = 0.830$, $SE = 0.042$, $z = 19.643$, $p < .001$, 95% CI [0.747, 0.913]). Collectively, these findings provided further evidence for the PXI's theoretical model and overall criterion validity.

5 DISCUSSION

We have presented results from an independent psychometric evaluation of the PXI using a large sample. The PXI is a promising scale for measuring the functional and psychosocial consequences of playing digital games [1]. However, independent validation of the scale was yet to be conducted [65]. For this reason, we set up a pre-registered online study and collected data from 1518 participants. With the collected data, we conducted various forms of psychometric quality analysis to evaluate the PXI. Results, in general, show that the PXI performs well regarding commonly used scale reliability and validity indicators, with good CFA model fits and satisfactory internal consistency values for all constructs except immersion. In addition, results mostly favored the convergent and discriminant validity of the PXI while further supporting the scale's criterion validity.

Results from the present study are comparable to those reported in the original work on the PXI [1] and for the German version of the scale [24]. Most of the sample initially used to develop the PXI consisted of students. In contrast, the German PXI was developed using volunteer participants recruited over mailing lists, online groups, and social media. Also, samples from previous work on the PXI consisted predominantly of young men, not reflecting the demographics

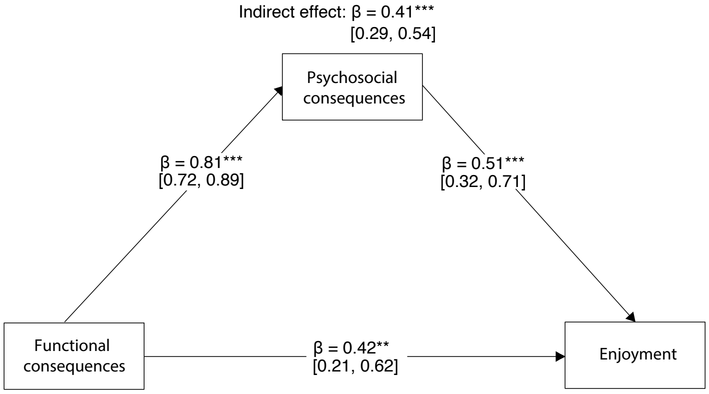


Fig. 1. Hybrid structural equation model displaying the relationship between the psychosocial consequences, the functional consequences, and the enjoyment items of the PXI. $*** = p < .001; ** = p < .01$.

of digital game players [20] and were below common sample size recommendations [40]. We recruited a large sample of participants using a crowd-sourcing platform, resulting in a more balanced sample regarding gender and age distribution. This is also important considering that about an equal share of men and women and a substantial amount of older adults play digital games [20]. Our findings thus highlight that the PXI retains consistent psychometric quality across various populations and performs just as well with a large sample of crowd-sourced participants and a more balanced sample concerning gender and age as with previously used samples.

However, the results also revealed particular challenges with the PXI, some comparable to previous work, which must be addressed. Concerning item analysis, multiple items exhibited low item variances. However, given the scale's performance for other psychometric analyses, we do not see these results as strong evidence against the PXI's quality. Potentially, the item variance was low for some items, such as enjoyment, because of the wording of the critical incident technique, which likely caused participants to pick games they generally enjoyed, resulting in primarily high ratings and thus low variance. Furthermore, four constructs had AVE values below the desired threshold, indicating room for improvement concerning convergent validity: mastery, immersion, challenge, and ease of control. This suggests that the items for these constructs are not as closely related as would be expected if they formed a common factor. While ease of control was the only construct that exhibited a sub-optimal AVE value in the original PXI paper, it barely met the threshold in the German PXI paper. Consequently, there appears to be room for improvement regarding the convergent validity of ease of control. For the constructs of mastery and challenge, past work reported no problematic results. It thus remains to be determined if similar challenges with these constructs will arise in future work or if they are unique to the present study. Given that the construct of immersion was conspicuous not only regarding the AVE but also in other analyses, we will return to it in the following subsection 5.1. Finally, regarding criterion validity, the constructs of progress feedback and goals and rules exhibited only weak correlations to the related construct of pragmatic quality. Given that these constructs correlated as expected in the original PXI validation, it is unclear whether our experimental setup caused the low correlations or whether these results were not due to the PXI but rather because of the construct of pragmatic quality measured with the AttrakDiff. The criterion validity of progress feedback and goals and rules, as well as the question of whether the AttrakDiff holds up in the context of digital games, could thus be further explored in future work.

5.1 Challenges of the immersion construct

Based on the present work's results, the PXI's immersion construct was negatively salient in several respects. This is especially concerning given the frequent measurement of the immersion construct in research employing the PXI. The challenges of immersion in our independent validation are varied. Concerning reliability, immersion was the only construct of the PXI with internal consistency values below the desired threshold, although just barely. Regarding convergent validity, immersion fell below the desired value of $AVE \geq .50$. Furthermore, immersion was the sole construct with MSV values smaller than AVE while exhibiting a lower square root of AVE than the inter-construct correlations with multiple other constructs, which speaks against the construct's discriminant validity, suggesting that the immersion items are too closely related to the other factors. While the original PXI paper reported no challenges with immersion, the German PXI also showed issues with the discriminant validity of immersion. Finally, the item immersion_1 showed conspicuous descriptive statistics deviating from the other items during item evaluation.

Immersion being a difficult construct to both define and measure is not a novel problem within PX research. Previous measurements, such as the Game Engagement Questionnaire [GEQ, 8] have already encountered difficulties when externally validated [7, 50]. Further, we can see the presence of circular definitions of immersion in the GEQ's developmental paper, i.e. the researchers explain immersion as being engaged in an activity, but immersion is also a construct in their engagement questionnaire.

> "Immersion is typically used to describe the experience of becoming engaged in the game-playing experience while retaining some awareness of one's surroundings." [8, p. 624]

The original development paper of the PXI offers a definition of immersion, which also appears to be tautological in nature;

> "A sense of immersion and cognitive absorption, experienced by the player" [1, p. 5]

By comparing the GEQ's and the PXI's definition of immersion, we find the GEQ treats absorption as its own construct. Following, the two operationalizations disagree on whether an individual would be aware of their surroundings or not when experiencing immersion. See the PXI's item immersion_1 "I was no longer aware of my surroundings while I was playing." in comparison to the GEQ's definition. This item was also answered with the greatest variance by our sample, with answers tending towards either end of the scale. This warrants further theoretical and methodological investigation of whether a lack of awareness of one's surroundings is an aspect of immersion or should constitute an alternative construct, such as absorption. Although recent literature has attempted to increase the clarity for the immersion construct [2] and other similar constructs concerning psychological absorption in a task [35], there remains more work to be done in regards to a more consistent definition of what immersion entails.

Additionally, according to MacKenzie [43], a good definition needs to be unambiguously distinguished from related constructs. However, it has remained difficult to differentiate immersion. In the case of the PXI, we are unable to clearly differentiate immersion from the constructs of meaning and enjoyment, which is evidenced by the results on divergent validity. As such, there is more theoretical work necessary to distinguish immersion from other constructs and subsequently achieve a more robust operationalization.

5.2 Model behind the PXI

One question we wanted to answer with the present study was which theoretical model should be employed when working with the PXI. As stated in the related work, the information provided by the authors, both in the original paper and on the PXI's website, provided no definitive suggestions on what exact theoretical model should be used for the PXI, if a model should include higher-order factors such as for the functional and psychosocial consequences, and how the suggested items for enjoyment would relate to the items of the PXI. Furthermore, our literature review of the usage of the PXI also showed different measurement models being used for the scale.

Based on the present results, especially from the CFAs, we found the most robust evidence for a ten-factor model with one factor per construct of the PXI. A simple ten-factor model exhibited a better model fit compared to more complex models that also considered higher-order factors for the psychosocial and functional consequences or an overall factor for PX. Consequently, we recommend that authors working with the PXI stick to a model with one factor per construct and do not form higher-order scores for the psychosocial and functional consequences or an overall PX factor. Such a ten-factor model is also in line with most of what the original authors themselves suggest, both in their paper and on the official website of the PXI. Finally, we also considered the enjoyment items the original authors suggested to be used alongside the PXI, but for which the psychometric quality still was to be investigated. Our results showed that the enjoyment items perform comparably well to the PXI items. At the same time, their inclusion in the scale and the resulting consideration of an enjoyment factor and an 11-factor model did not negatively affect the quality of the overall scale. Thus, these items can be used alongside the PXI in good conscience.

5.3 Weak evidence of a general player experience score

One common discrepancy between the PXI's theoretical foundation and its application in practice is the averaging of all items into one overall score of PX. As seen above, we find the statistical model does not lend itself to this interpretation of the scale, with a better model fit exhibited for those models that do not contain a higher-order factor of PX. However, some researchers average all responses to items of the PXI into one singular score. For example, one paper employed the PXI to score their game on reaching a certain amount of points out of the 90-point total score achievable in the PXI. While such a general PX score was not validated in the original paper, nor theoretically proposed by the original authors, we investigated such a model to see whether an average score would be appropriate. The results showed that the introduction of such a general PX factor into the model worsened the model fit compared to a ten-factor model without a higher-order factor. Given this finding, we caution both researchers and practitioners against using the PXI to measure the construct of PX and rather interpret the responses to the individual constructs with intention and care. Furthermore, we see two additional reasons that speak against a general PX score. First, digital games and other interactive media relevant to the PXI come in a wide variety of forms with many different goals. A certain digital game might not have been designed to provide ease of control and instead was designed for a particularly high level of difficulty. A low score in ease of control would thus not mean this is a design problem to be fixed. Second, there are no guidelines or cut-offs from the original authors as to what would constitute a satisfactory score, e.g. for enjoyment. Therefore, we can not recommend applying the PXI to determine whether a game has good or bad PX in a simplistic manner.

5.3.1 *Applicability of the PXI.*

Following, we find the strength of the PXI in comparing different games and, specifically, different versions of the same game in terms of their experiential quality. Indeed, the PXI provides a variety of relevant constructs along which player experiences can be compared and contrasted. For practitioners, the PXI can aid in the incremental development of games and testing for version improvements. This interpretation of the PXI's strengths is also in line with recommendations by the original authors. The applicability regarding comparison is further enabled through the use of the PXI bench, which has collected PXI data across different games and genres [26]. For researchers, we find the PXI useful to study its constructs as a dependent measure to compare between player experiences, similar to the applicability for practitioners. However, theoretical engagement with the constructs prior to measurement is still required, especially regarding constructs such as meaning or immersion, as they are not fully differentiated in theory or construct validity.

6 LIMITATIONS

As a first limitation, we did not have participants interact with a digital game but rather recall a memorable game using the critical incident technique well-established in HCI [e.g., 6, 62]. While this task is comparable to those used in other work on the PXI, we cannot exclude that the chosen experimental task influenced certain results, such as the low item variance for some items. Because participants were instructed to think of a game they recently played or remembered well, they presumably mainly chose games they liked well, which might have caused the low variance for some items, such as for enjoyment. This is also reflected in the skewed distribution of responses to most items in the present study. While also likely due to the wording of the critical incident technique, which probably caused participants to pick games they generally enjoyed, it does affect the generalizability of the present results to other contexts. More research is needed in this regard to investigate whether this is an issue generated by the research methodology used or if it is a general challenge affecting the applicability of the PXI. Furthermore, it is possible that participants could not remember the games very well, thus influencing their responses compared to actual interaction with a digital game and consequently the external validity of the reported findings. Hence, the psychometric quality of the PXI still needs to be further investigated after actual interactions with digital games. This approach would likely be closer to the scale's intended use compared to the critical incident technique, increasing the external validity of such results. Initial findings in this regard were already reported in the original work on the PXI [1], and just recently, Haider et al. [28] reported on a preliminary investigation on the miniPXI's potential to evaluate prototypes during game development. Second, we collected data using an online study setting. While this is comparable to the procedure used in past work on the PXI and allowed us to collect a sufficiently large sample needed to conduct certain analyses (e.g., CFA), results from data collected in a lab study might differ from the ones reported here.

One general limitation of the statistical analysis of construct validity is the influence of the chosen wording per item on the consistency of the subjects' responses. Maul [45] found that items of nonsensical expressions, but with consistent wording, would still show acceptable fit in factor analysis. Indeed, the PXI is constructed of items that display consistent wording within their respective sub-factor. For example, all items relating to autonomy begin with the wording "I felt [...]", and two of the three of them end with "[...] I wanted to play this game". These choices regarding wording have an influence on the statistical validation process. However, we cannot account for the magnitude of this influence. Furthermore, these consistent wordings can also lead to complex sub-factors, such as immersion, being limited in the breadth with which they construct this experience. Statistical validation can not account for content validation, and therefore, we fundamentally can not determine whether the items presented in the PXI genuinely reflect the experiences they wish to measure [3]. These challenges to construct validity are as old as the method itself [5]. As such, we aim to provide a fair and balanced interpretation of our work and the general findings on the evidence of validity for the PXI rather than a definitive endorsement for the measurement.

7 FUTURE WORK

As mentioned above, the present study worked with self-reported experiences collected using an online survey. While this procedure closely matches past work on the PXI, it comes with certain limitations. Consequently, it would be interesting to see how the PXI performs in a lab study setting more comparable to a GUR evaluation in the industry. Initial evidence for the PXI's performance in such a setting was reported in the original paper, indicating that the PXI had configural but not metrical invariance between an online study collecting recalled experience data and ratings from experimental investigations or play tests relying on immediate recall after playing [1]. While out of scope for the present work, gathering additional data on the scale's performance in such a setting would be intriguing for future independent validations of the PXI to see how the quality of the scale compares across settings. We further see an opportunity for future research to investigate if the PXI can differentiate between different versions of the same game, for example, after improvements and changes have been made, and if changes made to particular aspects of the game are also reflected in corresponding ratings of the PXI (e.g., audiovisual design changes resulting in a different score for the PXI's audiovisual appeal rating). Such efforts could also be used to examine the PXI's criterion validity, for example, by showing that the experimental manipulation of certain game design elements leads to changes in the respective constructs of the scale. Furthermore, our results were strongly positively skewed. Therefore, we find a potential for future work to investigate whether the PXI can differentiate between different player experiences, such as comparing particularly positive experiences to mainly negative experiences with the same or other digital games. In addition, the present work did not investigate the psychometric quality of the 11-item short version of the PXI, the miniPXI [27]. While beyond the scope of the current study, re-investigating the quality of this scale version poses an additional opportunity for future work. Finally, while immersion was conspicuous in our sample, previous research on the PXI has not reported on comparable problems for this specific construct but for others (e.g., low AVE for ease of control in the original PXI paper). To deepen the understanding of the PXI's psychometric quality and the stability of its constructs across various settings and populations, researchers who use the PXI should, if the sample size permits it, investigate the psychometric quality again or otherwise provide their data so that future validation studies could do such analyses.

8 CONCLUSION

The present paper reported on a large-sample independent validation of the PXI, a scale measuring the psychosocial and functional consequences of playing digital games. In a pre-registered online study, 1518 participants rated a recent or memorable digital game using all items of the PXI and a selection of related scales. Results showed that the PXI performs well, with common indicators of psychometric quality delivering acceptable to excellent results. Furthermore, results showed that the enjoyment items proposed to be used alongside the PXI are also of good quality and can thus be employed alongside the scale. However, immersion was identified and discussed as a challenging construct as it could not be clearly distinguished from meaning or enjoyment. Finally, results demonstrated that the theoretical model behind the PXI is best understood as consisting of one individual factor per construct of the PXI, without any higher-order factors. Overall, the results demonstrated that researchers can confidently use the PXI in their studies.

9 DATA AVAILABILITY STATEMENT

The pre-registration (https://osf.io/buq5t/?view_only=5d69aed003e94ac5b04820c33fdb101a) and supplementary materials (https://osf.io/8xuhr/?view_only=46df6e3da1b44824ae5c0bbfd5ad695a) for this study are available on OSF.

10 FUNDING AND DECLARATION OF CONFLICTING INTERESTS

This work is financed entirely by our research group, as we received no additional funding. The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (2020), 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

[2] Lena Fanya Aeschbach, Klaus Opwis, and Florian Brühlmann. 2022. Breaking immersion: A theoretical framework of alienated play to facilitate critical reflection on interactive media. *Frontiers in Virtual Reality* 3 (2022), 1–14. https://doi.org/10.3389/frvir.2022.846490

[3] Lena Fanya Aeschbach, Sebastian A. C. Perrig, Lorena Weder, Klaus Opwis, and Florian Brühlmann. 2021. Transparency in Measurement Reporting: A Systematic Literature Review of CHI PLAY. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 233 (10 2021), 21 pages. https://doi.org/10.1145/3474660

[4] Hirotugu Akaike. 1987. Factor analysis and AIC. *Psychometrika* 52 (1987), 317–332. https://doi.org/10.1007/BF02294359

[5] Harold P. Bechtoldt. 1959. Construct validity: A Critique. *American Psychologist* 14, 10 (1959), 619–629. https://doi.org/10.1037/h0040359

[6] Julia Ayumi Bopp, Elisa D. Mekler, and Klaus Opwis. 2016. Negative Emotion, Positive Experience? Emotionally Moving Moments in Digital Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2996–3006. https://doi.org/10.1145/2858036.2858227

[7] John J. Bosley. 2013. Gauging Engagement in Video Games: Does Game Violence Relate to Player Behavior? Report on a Study. *Interacting with Computers* 25, 4 (July 2013), 284–286. https://doi.org/10.1093/iwc/iwt006

[8] Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634. https://doi.org/10.1016/j.jesp.2009.02.016

[9] Timothy A. Brown. 2015. *Confirmatory factor analysis for applied research* (2 ed.). The Guilford Press, New York, NY, USA.

[10] Florian Brühlmann and Elisa D. Mekler. 2018. Surveys in Games User Research. In *Games User Research*. Oxford University Press, Oxford, UK. https://doi.org/10.1093/oso/9780198794844.003.0009

[11] Florian Brühlmann, Serge Petralito, Lena Aeschbach, and Klaus Opwis. 2020. The Quality of Data Collected Online: An Investigation of Careless Responding in a Crowdsourced Sample. *Methods in Psychology* 2 (2020), 100022. https://doi.org/10.1016/j.metip.2020.100022

[12] Peter Cabrera-Nguyen. 2010. Author Guidelines for Reporting Scale Development and Validation Results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research* 1, 2 (2010), 99–103. https://doi.org/10.5243/jsswr.2010.8

[13] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334. https://doi.org/10.1007/BF02310555

[14] Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (2016), 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

[15] Edward L. Deci. 1975. *Intrinsic Motivation*. Plenum Press, New York, NY, USA. https://doi.org/10.1007/978-1-4613-4446-9

[16] Edward L. Deci and Richard M. Ryan. 1980. The Empirical Exploration of Intrinsic Motivational Processes. In *Advances in Experimental Social Psychology*, Leonard Berkowitz (Ed.). Vol. 13. Academic Press, New York, NY, USA, 39–80. https://doi.org/10.1016/S0065-2601(08)60130-6

[17] Edward L. Deci and Richard M. Ryan. 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11, 4 (2000), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

[18] Robert F. DeVellis. 2017. *Scale development: Theory and applications* (4 ed.). SAGE publications, Inc., Thousand Oaks, CA, USA.

[19] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* 18, 3 (2023), e0279720. https://doi.org/10.1371/journal.pone.0279720

[20] Maeve Duggan. 2015. *Gaming and Gamers*. Pew Research Center. Retrieved August 24, 2023 from https://www.pewresearch.org/internet/2015/12/15/gaming-and-gamers/

[21] Thomas J. Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 3 (2014), 399–412. https://doi.org/10.1111/bjop.12046

[22] Mike Furr. 2011. *Scale construction and psychometrics for social and personality psychology* (1 ed.). SAGE publications, Ltd., London, UK.

[23] Darren George and Paul Mallery. 2019. *IBM SPSS statistics 26 step by step: A simple guide and reference* (16 ed.). Routledge, New York, NY, USA. https://doi.org/10.4324/9780429056765

[24] Linda Graf, Maximilian Altmeyer, Katharina Emmerich, Marc Herrlich, Andrey Krekhov, and Katta Spiel. 2022. Development and Validation of a German Version of the Player Experience Inventory (PXI). In *Proceedings of Mensch Und Computer 2022* (Darmstadt, Germany) *(MuC '22)*. Association for Computing Machinery, New York, NY, USA, 265–275. https://doi.org/10.1145/3543758.3543763

[25] Jonathan Gutman. 1982. A Means-End Chain Model Based on Consumer Categorization Processes. *Journal of Marketing* 46, 2 (1982), 60–72. https://doi.org/10.1177/002224298204600207

[26] Aqeel Haider, Kathrin Gerling, and Vero Vanden Abeele. 2020. The Player Experience Inventory Bench: Providing Games User Researchers Actionable Insight into Player Experiences. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 248–252. https://doi.org/10.1145/3383668.3419898

[27] Aqeel Haider, Casper Harteveld, Daniel Johnson, Max V. Birk, Regan L. Mandryk, Magy Seif El-Nasr, Lennart E. Nacke, Kathrin Gerling, and Vero Vanden Abeele. 2022. MiniPXI: Development and Validation of an Eleven-Item Measure of the Player Experience Inventory. *Proc. ACM Hum.-Comput. Interact.* 6, CHI PLAY, Article 244 (10 2022), 26 pages. https://doi.org/10.1145/3549507

[28] Aqeel Haider, Günter Wallner, Kathrin Gerling, and Vero Vanden Abeele. 2023. Preliminary Study of the Performance of the MiniPXI When Measuring Player Experience throughout Game Development. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Stratford, ON, Canada) *(CHI PLAY Companion '23)*. Association for Computing Machinery, New York, NY, USA, 56–62. https://doi.org/10.1145/3573382.3616076

[29] Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. 2010. *Multivariate Data Analysis* (7 ed.). Prentice Hall, Hoboken, NJ, USA.

[30] Marc Hassenzahl. 2004. The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction* 19, 4 (2004), 319–349. https://doi.org/10.1207/s15327051hci1904_2

[31] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003: Interaktion in Bewegung*, Gerd Szwillus and Jürgen Ziegler (Eds.). Vieweg+Teubner Verlag, Wiesbaden, Germany, 187–196. https://doi.org/10.1007/978-3-322-80058-9_19

[32] N. Henze and B. Zirkler. 1990. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods* 19, 10 (1990), 3595–3617. https://doi.org/10.1080/03610929008830400

[33] Daire Hooper, Joseph Coughlan, and Michael R. Mullen. 2008. Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods* 6, 1 (2008), 53–60.

[34] Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1 (1999), 1–55. https://doi.org/10.1080/10705519909540118

[35] Kyros Jalife, Casper Harteveld, and Christoffer Holmgård. 2021. From Flow to Fuse: A Cognitive Perspective. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 256 (10 2021), 30 pages. https://doi.org/10.1145/3474683

[36] Dominik Kayser, Sebastian A. C. Perrig, and Florian Brühlmann. 2021. Measuring Players' Experience of Need Satisfaction in Digital Games: An Analysis of the Factor Structure of the UPEQ. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Austria) *(CHI PLAY '21)*. Association for Computing Machinery, New York, NY, USA, 158–162. https://doi.org/10.1145/3450337.3483499

[37] Ken Kelley. 2007. Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software* 20, 8 (2007), 1–24. https://doi.org/10.18637/jss.v020.i08

[38] Ken Kelley. 2007. Methods for the Behavioral, Educational, and Social Sciences: An R package. *Behavior Research Methods* 39 (2007), 979–984. https://doi.org/10.3758/BF03192993

[39] T. L. Kelley. 1927. *Interpretation of educational measurements* (1 ed.). World Book Co., Oxford, England.

[40] Rex B. Kline. 2016. *Principles and practice of structural equation modeling* (4 ed.). Guilford Press, New York, NY, USA.

[41] Jennifer G. La Guardia, Richard M. Ryan, Charles E. Couchman, and Edward L. Deci. 2000. Within-person variation in security of attachment: a self-determination theory perspective on attachment, need fulfillment, and well-being. *Journal of personality and social psychology* 79, 3 (2000), 367–384. https://doi.org/10.1037/0022-3514.79.3.367

[42] Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. 2018. Systematic Review and Validation of the Game Experience Questionnaire (GEQ) - Implications for Citation and Reporting Practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 257–270. https://doi.org/10.1145/3242671.3242683

[43] Scott B. MacKenzie. 2003. The Dangers of Poor Construct Conceptualization. *Journal of the Academy of Marketing Science* 31, 3 (2003), 323–326. https://doi.org/10.1177/0092070303031003011

[44] Kantilal Vardichand Mardia. 1970. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 3 (1970), 519–530. https://doi.org/10.1093/biomet/57.3.519

[45] Andrew Maul. 2017. Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives* 15, 2 (2017), 51–69. https://doi.org/10.1080/15366367.2017.1348108

[46] Roderick P. McDonald. 1999. *Test theory: A unified treatment* (1 ed.). Psychology Press, New York, NY, USA. https://doi.org/10.4324/9781410601087

[47] Adam W. Meade and S. Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437–455. https://doi.org/10.1037/a0028085

[48] Elisa D. Mekler, Julia Ayumi Bopp, Alexandre N. Tuch, and Klaus Opwis. 2014. A Systematic Review of Quantitative Studies on the Enjoyment of Digital Entertainment Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 927–936. https://doi.org/10.1145/2556288.2557078

[49] Zgjim Memeti, Florian Brühlmann, and Sebastian A. C. Perrig. 2022. LoL, Why Do You Even Play? Validating the Motives for Online Gaming Questionnaire in the Context of League of Legends. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play* (Bremen, Germany) *(CHI PLAY '22)*. Association for Computing Machinery, New York, NY, USA, 81–86. https://doi.org/10.1145/3505270.3558350

[50] Kent L. Norman. 2013. GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers* 25, 4 (2013), 278–283. https://doi.org/10.1093/iwc/iwt009

[51] Jum C. Nunnally. 1978. *Psychometric Theory* (2 ed.). Mcgraw hill book company, New York, NY, USA.

[52] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54 (2022), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

[53] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[54] Thomas J. Reynolds and Jonathan Gutman. 2001. *Laddering theory, method, analysis, and interpretation.* Lawrence Erlbaum Associates Publisher, Mahwah, NJ, US, 25–62.

[55] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. https://doi.org/10.18637/jss.v048.i02

[56] Richard M. Ryan and Edward L. Deci. 2000. *Intrinsic Motivation Inventory (IMI).* Center for Self-Determination Theory. Retrieved August 24, 2023 from https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

[57] Richard M. Ryan and Edward L. Deci. 2001. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology* 52, 1 (2001), 141–166. https://doi.org/10.1146/annurev.psych.52.1.141

[58] Richard M. Ryan, Valerie Mims, and Richard Koestner. 1983. Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of personality and Social Psychology* 45, 4 (1983), 736–750. https://doi.org/10.1037/0022-3514.45.4.736

[59] Richard M Ryan, C. Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* 30 (2006), 344–360. https://doi.org/10.1007/s11031-006-9051-8

[60] Albert Satorra and Peter M. Bentler. 2001. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66, 4 (2001), 507–514. https://doi.org/10.1007/BF02296192

[61] Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464. http://www.jstor.org/stable/2958889

[62] Mirjam Seckler, Silvia Heinz, Seamus Forde, Alexandre N. Tuch, and Klaus Opwis. 2015. Trust and distrust on the web: User experiences and website characteristics. *Computers in Human Behavior* 45 (2015), 39–50. https://doi.org/10.1016/j.chb.2014.11.064

[63] Holger Steinmetz. 2015. *Lineare Strukturgleichungsmodelle. Eine Einführung mit R [Linear Structural Equation Modeling. An introduction with R]* (2 ed.). Rainer Hampp Verlag, Mering; München, Germany.

[64] Barbara G. Tabachnick and Linda S. Fidell. 2007. *Using multivariate statistics* (5 ed.). Allyn & Bacon, Inc., Needham Heights, MA, USA.

[65] April Tyack and Peta Wyeth. 2021. "The Small Decisions Are What Makes It Interesting": Autonomy, Control, and Restoration in Player Experience. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 282 (10 2021), 26 pages. https://doi.org/10.1145/3474709

[66] Vero Vanden Abeele, Lennart E. Nacke, Elisa D. Mekler, and Daniel Johnson. 2016. Design and Preliminary Validation of The Player Experience Inventory. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (Austin, Texas, USA) *(CHI PLAY Companion '16)*. Association for Computing Machinery, New York, NY, USA, 335–341. https://doi.org/10.1145/2968120.2987744

[67] Nick von Felten, Florian Brühlmann, and Sebastian A. C. Perrig. 2022. Independent Validation of the Video Game Dispositional Flow Scale With League of Legends Players. In *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play* (Bremen, Germany) *(CHI PLAY '22)*. Association for Computing Machinery, New York, NY, USA, 44–50. https://doi.org/10.1145/3505270.3558351

[68] Peter Vorderer, Christoph Klimmt, and Ute Ritterfeld. 2004. Enjoyment: At the Heart of Media Entertainment. *Communication Theory* 14, 4 (01 2004), 388–408. https://doi.org/10.1111/j.1468-2885.2004.tb00321.x

[69] Christina Werner and Karin Schermelleh-Engel. 2010. *Deciding Between Competing Models: Chi-Square Difference Tests*. Goethe University, Frankfurt. Retrieved November 24, 2023 from https://www.researchgate.net/publication/241278052_Deciding_Between_Competing_Models_Chi-Square_Difference_Tests

[70] Robert W. White. 1959. Motivation reconsidered: the concept of competence. *Psychological review* 66, 5 (1959), 297–333. https://doi.org/10.1037/h0040934