

**Universität
Basel**

Fakultät für
Psychologie



Understanding and Improving Subjective Measures in Human-Computer Interaction

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

Florian Brühlmann

aus Aarau

Basel, 2018



Universität
Basel

Fakultät für
Psychologie



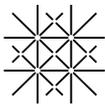
Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Klaus Opwis

Dr. Javier Bargas-Avila

Datum des Doktoratsexamen:

DekanIn der Fakultät für Psychologie



Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass 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:

- Brühlmann, F., Vollenwyder, B., Opwis, K., & Mekler, E. D. (2018). Measuring the "why" of interaction: Development and validation of the user motivation inventory (UMI). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. (pp. 106:1–106:13). New York, NY, USA: ACM. doi: 10.1145/3173574.3173680
- Bargas-Avila, J.A. & Brühlmann, F. (2016). Measuring user rated language quality: Development and validation of the user interface language quality survey (LQS). *International Journal of Human-Computer Studies*, 86, 1-10. doi: 10.1016/j.ijhcs.2015.08.010
- Brühlmann, F., Petralito, S., Rieser, D. C., Aeschbach, L. F., & Opwis, K. (2018). TrustDiff: *Development and validation of a semantic differential for user trust on the web*. Manuscript submitted for publication.
- Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2018). *Half of the participants in online surveys respond carelessly: An investigation of data quality in crowdsourced samples*. Manuscript submitted for publication.

SUBJECTIVE MEASURES IN HCI	3
Contents	
Erklärung zur wissenschaftlichen Lauterkeit	2
Abstract	4
Introduction	5
Issues with questionnaires and theories in HCI	7
User motivation	8
User interface language quality	10
User trust	11
Careless responding and online research	12
Summary of the Manuscripts	15
Manuscript 1: Measuring the "Why" of Interaction: Development and Validation of the User Motivation Inventory (UMI)	18
Manuscript 2: Measuring user rated language quality: Development and Validation of the user interface Language Quality Survey (LQS)	23
Manuscript 3: TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web	26
Manuscript 4: Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples	31
General Discussion	37
Validity and validation	37
Theory and measurement	38
Careless responding and online research	41
Limitations and future directions	42
Conclusion	45
References	45
Acknowledgements	56
Curriculum Vitae	57
Appendix	58

Abstract

In Human-Computer Interaction (HCI), research has shifted from a focus on usability and performance towards the holistic notion of User Experience (UX). Research into UX places special emphasis on concepts from psychology, such as emotion, trust, and motivation. Under this paradigm, elaborate methods to capture the richness and diversity of subjective experiences are needed. Although psychology offers a long-standing tradition of developing self-reported scales, it is currently undergoing radical changes in research and reporting practice. Hence, UX research is facing several challenges, such as the widespread use of ad-hoc questionnaires with unknown or unsatisfactory psychometric properties, or a lack of replication and transparency. Therefore, this thesis contributes to several gaps in the research by developing and validating self-reported scales in the domain of user motivation (manuscript 1), perceived user interface language quality (manuscript 2), and user trust (manuscript 3). Furthermore, issues of online research and practical considerations to ensure data quality are empirically examined (manuscript 4). Overall, this thesis provides well-documented templates for scale development, and may help improve scientific rigor in HCI.

Introduction

In the last decade, research on Human-Computer Interaction (HCI) has moved from a focus on usability and performance towards the more holistic view of user experience (UX). Moreover, UX research aims to go beyond pragmatic-instrumental aspects of technology use (to what extent a technology helps to achieve a goal) and tries to understand how non-instrumental and hedonic aspects of technologies (such as having fun and self-expression) can contribute to the overall perception of product quality (Hassenzahl & Tractinsky, 2006). Today, digital technologies are no longer expected to be simply intuitive and easily learned, but should also enrich our lives by providing meaningful and aesthetic experiences. Hassenzahl and Tractinsky (2006)'s understanding of UX emphasizes its situatedness and temporality, which presents unique challenges in evaluation and measurement. New models and research methods need to be developed to capture different aspects of the subjective user experience holistically. However, subjective experiences are inherently difficult to capture in a reliable, objective, and valid way (DeVellis, 2016). Various research areas in psychology have a long-standing tradition in developing measures to study subjective experience of affect, cognition, and evaluation. This is not surprising, because measuring and understanding subjective phenomena is one of the pillars of modern psychology. Derived from its origins in intelligence tests and assessments, measures of a large variety of constructs such as personality (O. P. John & Srivastava, 1999), depression (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), work motivation (Gagné et al., 2015) or life satisfaction (Diener, Emmons, Larsen, & Griffin, 1985) have been developed and applied in both research and practice.

Presently, psychology is undergoing large and radical changes in research practice (Hesse, 2018). These changes originate from very unlikely results published in one of the top psychology journals, the *Journal for Personality and Social Psychology* (2011). The paper by Bem (2011) reported evidence for para-psychological phenomena in a very convincing way, which were then discussed controversially in the community. The debate was mainly concerned with the fact that such a bold claim could be published without independent replication and transparency in materials and statistical analyses (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). In the same year, Simmons, Nelson, and Simonsohn (2011) demonstrated how a few decisions about data collection and analysis could alter results drastically, presenting any difference as significant. Four years later, in an international collaborative effort, 100 contemporary psychology experiments were tested (according to their replicability) by the Open Science Collaboration. Results indicated that depending on the measure, only between 36% to 47% of the studies successfully replicated (Open Science Collaboration, 2015). This phase was termed the *replication crisis* (Pashler & Wagenmakers, 2012). The causes of these failed replications are commonly perceived as how

psychologists conducted research, and how research is incentivized: Surprising and statistically significant results were published with ease, which incentivized researchers to engage in questionable research practices, such as only reporting parts of an experiment (or changing hypotheses) after the results were known (L. K. John, Loewenstein, & Prelec, 2012). Researchers are often not aware that these decisions can greatly influence Type-1 errors and lead to false, non-replicable conclusions (L. K. John et al., 2012). This effect is reinforced by the frequency of low-powered studies, which are especially prone to these issues (Ioannidis, 2005). In recent years, several strategies to counter this problem have been developed, including preregistration of studies, open data and materials, encouragement of replications, publishing null findings, and large, high-powered international collaborations (e.g., Buttrick et al., 2018; Nosek et al., 2015).

Although HCI has always been heavily influenced by psychology (Dix, 2017), the lack of replication in the HCI community has been debated, even before the replication crisis in psychology gained traction (e.g., Wilson et al., 2011). In 2012, Kaptein and Robertson (2012) introduced several issues discussed in psychology related to HCI, such as low-powered studies and misinterpretation of p-values. Despite great interest in these topics, replications remained rare (Hornbæk, Sander, Bargas-Avila, & Simonsen, 2014). Recently, under increased community interest, new initiatives have been implemented to improve research practice in HCI (e.g., Kay, Haroz, Guha, & Dragicevic, 2016; Kay, Haroz, Guha, Dragicevic, & Wacharamanotham, 2017). This time, the focus is broader, with discussions including replications, research practices in general, and even how HCI might contribute to the development of tools that enable researchers to make less questionable decisions (Chuang & Pfeil, 2018; Cockburn, Gutwin, & Dix, 2018; Echtler & Häussler, 2018). Apart from incentives and research practices, the validity and replicability of research and theory building in empirical science depends heavily on measured data. Data captured with various measures is one of the most essential sources for understanding relationships, causes, and effects, and explaining phenomena. However, even a robust study methodology fails when noisy data, unreliable measures, or systematic biases is introduced by data collection. Such studies will be more difficult to replicate, and might lead to false decisions (Loken & Gelman, 2017). Hence, proper operationalization, precise measurement, and data quality are essential factors for reliable and valid conclusions. Measurement of phenomena is vital for discovering actual causal mechanisms and theory development (Bringmann & Eronen, 2016). Many of these aspects of high-quality research still need improving in HCI. Besides replication, HCI lacks theory (Liu et al., 2014; Oulasvirta & Hornbæk, 2016), and measures that meet psychometric standards (Bargas-Avila & Hornbæk, 2011). For instance, there is an ongoing controversy about the Game Experience Questionnaire (GEQ; Poels, de Kort, & Ijsselstein, 2007), which is one of the most widely adopted scales in the growing field of

Player Experience research (Brühlmann & Mekler, 2018; Law, Brühlmann, & Mekler, 2018). However, the GEQ has repeatedly failed the criteria of structural validity (Brühlmann & Schmid, 2015; D. Johnson, Gardner, & Perry, 2018; Law et al., 2018). To advance research in the field of HCI, freely available and well-studied questionnaires that can be applied to a wide range of products are helpful. Accordingly, this thesis contributes to three areas of HCI research (through the development of reliable and valid questionnaires), and to the study of data quality in online research.

In the context of these fundamental issues in HCI research, this thesis is concerned with many of these issues either directly or indirectly. Manuscript 1 reports on the development and validation of the User Motivation Inventory (UMI). It follows the principles of openness and transparency, as well as a theory-grounded approach that complies with best practices in questionnaire development. Manuscript 2 reports a bottom-up approach in development and validation of the user interface language quality survey (LQS). Translation quality is highly relevant in the development of products for a global market. Manuscript 3 concerns the development and validation of a semantic differential that measures user trust on the web. Semantic differentials are influential in UX research, requiring special considerations (Verhagen, van Den Hooff, & Meents, 2015). Complementing the first three studies, the fourth manuscript concerns methodological issues of data quality in online studies. Together, these manuscripts encourage thinking more clearly about measures, thorough test measures with modern statistical methods, and the employment of checks to ensure high-quality data.

Issues with questionnaires and theories in HCI

From its inception, HCI research has been influenced by various disciplines such as computer science, psychology, ergonomics, and social science (Dix, 2017). As outlined in the introduction, UX focuses more on the experiential aspects of interaction, such as emotions, aesthetics, and motives (Hassenzahl & Tractinsky, 2006). This shift of focus concerns aspects of experiences such as wellbeing, hedonic, and eudaimonic motives (Mekler & Hornbæk, 2016), utilizing concepts of positive psychology (Calvo & Peters, 2012). Despite encouraging, successful efforts to integrate concepts of psychology into HCI, a recent analysis of the proceedings of the CHI conference on Human-Computer Interaction (the most influential venue in HCI) indicated the field is lacking motor-themes and well-defined, influential theories (Liu et al., 2014). It appears that research in HCI is highly fragmented, and in a situation that Liu et al. (2014) describe as “when a new technology comes along it seems that researchers start from scratch leading to relatively isolated research themes” (p. 3560). One of the reasons for this might be the inherently interdisciplinary research approach, and its focus on new, emerging technologies. Another reason could be that there

are several different understandings of the central concepts, such as interaction (Hornbæk & Oulasvirta, 2017). Without a common language in the foundation of science (Rosenberg, 2011), building and testing overarching theories becomes challenging. The translation of theories and models into measurable, quantitative entities – *operationalizing* – is one of these central concepts of empirical science. Therefore, it is not surprising that Bargas-Avila and Hornbæk (2011) found that most studies in the HCI community employ ad-hoc questionnaires with little or no examination of psychometric properties. Without commonly accepted definitions of concepts, the operational definition of constructs in ad-hoc questionnaires follows the subjective understanding of the researcher and the context of the study. While such operationalizations may be appropriate in the specific context of one study, it hinders both wider generalization of findings beyond the study, and aggregation of evidence in meta-analyses. In the following sections, this challenge will be illustrated with three concrete applications.

User motivation

Motivation is a fundamental concept in our lives, driving intentional behavior. The reasons why people engage with an interactive technology affects how they use, perceive, and evaluate that technology, and the experience they had. However, very little is known about how motivation affects technology use and user experiences. Instead, different modes of use and context of user experiences have been studied by several authors in recent years (e.g., Deterding, 2016; Hassenzahl & Ullrich, 2007; Rozendaal, Keyson, & de Ridder, 2007; van Schaik & Ling, 2009). For instance, Mekler and Hornbæk (2016) described how user experiences with technology varied when pursued for eudaimonic (such as developing personal potential) or hedonic (such as pleasure) reasons. Eudaimonically motivated users experienced more need fulfillment, positive effects, and meaningful experiences compared to hedonically motivated users. Self-determination theory (SDT) is a promising theoretical framework for understanding how different motivations may influence technological experiences. It stems from positive psychology, and has already been successfully applied in various areas of user experience research (e.g., Deterding, 2016). Further, parts of the theory have even been integrated into user experience models (e.g., need satisfaction in Hassenzahl, Diefenbach, & Göritz, 2010). Self-determination theory describes motivational states and processes, and how they are shaped by individuals and social context (Deci & Ryan, 2000). A central postulate of SDT is that people experience varying degrees of basic psychological need satisfaction when they pursue an activity. The three basic psychological needs in SDT are for autonomy, competence, and relatedness. In SDT, need satisfaction is an outcome of goal attainment (the “what”), and the extent to which a certain activity supports need satisfaction is dependent on the underlying motivational regulation (the

Some aspects of SDT have been studied extensively in specific fields of HCI, such as need satisfaction in player experiences (Birk, Atkins, Bowey, & Mandryk, 2016; Deterding, 2016; Ryan, Rigby, & Przybylski, 2006) and experiences with technology (Hassenzahl et al., 2010; Hornbæk & Hertzum, 2017). However, research on OIT is scarce, and there is no research on motivation and technology use, which is rooted in SDT. Part of the reason for this research gap might be because there was no measuring instrument available. However, a better understanding of user motivation (the “why” of interaction) is imperative. Hence, the development of the User Motivation Inventory (UMI), a scale measuring motivational regulation based on OIT, which is the topic of manuscript 1 (Brühlmann, Vollenwyder, Opwis, & Mekler, 2018).

User interface language quality

The applied nature of HCI and UX research also creates problems such as context applicability, face validity, and efficiency of measures. When software is launched in a global market, it is vital to ensure that translation from the original user interface into other languages (localization) is of high quality. Most of the information in user interfaces is conveyed through text. Even graphical user interfaces rely heavily on language to communicate with users, and the text used to describe elements of navigation or the functionality of buttons varies between cultures and regions. For instance, informal text in user interfaces could be appropriate for the US but not in other cultures. Therefore, it is important to consider the correctness of translation and language, and the style and tone aspects of a specific culture. Translating user interface text has further specific challenges, such as *word sense disambiguation* (Muntés Mulero, Paladini Adell, España Bonet, & Màrquez Villodre, 2012). For example, the word “access” can represent “you have access” (a label) or “you can request access” (as a button) (Leiva & Alabau, 2014). Additionally, translating dates, genders, or prepositions without context frequently poses problems (Muntés Mulero et al., 2012). Mis-translations can affect user experiences negatively, and could result in lower trustworthiness, brand perception, acceptance, and perceived usefulness of a website (Sun, 2001). Therefore, it is important for products in multiple languages to monitor translation quality adequately.

Schriver (1989) describes three different classes of text quality evaluation: text-focused, expert-judgment-focused, and reader-focused. Text-focused evaluation includes automated methods, such as readability formulae (e.g., Fry, 1968; Kincaid, Fishburne Jr, Rogers, & Chissom, 1975) and are less suited for capturing contextual meanings of user interface text. Hence, reader- or expert-focused evaluation methods are more appropriate in the context of user interface translation. It has been demonstrated that expert evaluations increase the quality of interface text (Schriver, 1989), but have major limitations in terms

of time and resource constraints. In this situation, it might be more efficient to identify problems with reader-focused methods of text evaluation, such as through user surveys. These methods provide an initial test and help to prioritize expert evaluations of different languages accordingly. However, prior to the publication of manuscript 2, there was no readily applicable and validated measure of user perception of interface language quality. Therefore, it was decided to develop and validate a user interface language quality survey (LQS; Bargas-Avila & Brühlmann, 2016). The aim was to facilitate feedback for researchers and practitioners about the text quality of user interfaces; thus, enabling focused quality improvement efforts. Hence, the bottom-up scale development of LQs and user interface language quality is the topic of manuscript 2.

User trust

Trust was found to be one of the most important factors affecting the success of online transactions (Jarvenpaa, Tractinsky, & Saarinen, 1999; Schlosser, White, & Lloyd, 2006), and is crucial when users act under uncertainty (Casaló, Flavián, & Guinalú, 2007). Various academic fields study trust in different contexts (e.g. Driscoll, 1978; Moorman, Deshpande, & Zaltman, 1993; Rotter, 1967); therefore, there is no universally applicable definition. In recent years, trust in online contexts has been examined from various perspectives with different measures (Bhattacharjee, 2002; Cho, 2006; Flavián, Guinalú, & Gurrea, 2006; Gefen, 2002; McKnight, Choudhury, & Kacmar, 2002b). However, there is still no common, validated, reliable, and versatile measure (Kim & Peterson, 2017). Additionally, many measures of user trust have been tailored to specific contexts or websites (e.g. McKnight et al., 2002b). When researchers want to apply these methods in new contexts, they will need to rephrase items, possibly losing validity and reliability of the scale. Additionally, the scale developed by Flavián et al. (2006), which has been used in several studies (e.g. Seckler, Heinz, Forde, Tuch, & Opwis, 2015) was originally developed and validated in the Spanish language. Thus, it appears important to develop a scale that measures trust in various contexts of online shopping, and includes items that are easy to translate into different languages. Recent literature agrees that trust is a multidimensional construct composed of three different facets: benevolence, competence, and integrity (e.g., Bhattacharjee, 2002; Chen & Dhillon, 2003; Flavián et al., 2006; Gefen, 2002; Mayer, Davis, & Schoorman, 1995; McKnight et al., 2002b). These facets are defined as follows: Benevolence is defined as believing the other party is interested in their welfare (or a mutually beneficial relationship), and there is no intention of opportunistic behavior. Integrity (or honesty) is the belief that the other party is sincere and fulfills its promises. Competence describes the belief that the other party has the resources and capabilities needed for the successful completion of the transaction (Casaló et al., 2007). These three constructs have often been measured with

adapted questionnaires that use context-specific items such as “Do you agree that this C2C [Customer-to-Customer] platform solves a security problem or stops a fraudulent behavior?” (Lu, Wang, & Hayes, 2012). Therefore, we decided to develop a new measure for trust that does not rely on such specific characteristics or statements, termed the TrustDiff (Brühlmann, Petralito, Rieser, Aeschbach, & Opwis, 2018). The format of a semantic differential scale was chosen because it has several advantages over Likert-type scales (Verhagen et al., 2015). For instance, semantic differentials allow respondents to express opinions more fully than Likert-type scales, because disagreeing with an item on an agreement-scale does not necessarily mean agreeing with the opposite statement. Semantic differentials have also been found to be less prone to acquiescence bias (Friborg, Martinussen, & Rosenvinge, 2006), more robust, more reliable (Hawkins, Albaum, & Best, 1974; Wirtz & Lee, 2003), and under certain circumstances more valid (Van Auken & Barry, 1995). Semantic differential scales are especially suitable for efficiently measuring complex constructs (Chin, Johnson, & Schwarz, 2008; Verhagen et al., 2015). Investigation of these models will contribute to practice with a versatile and validated scale, further inform theory, and allow researchers to refine the three-factor model. Therefore, the development of a model-driven semantic differential scale for measuring user trust is the topic of manuscript 3.

Careless responding and online research

Online surveys have become a standard method of data collection in various fields such as psychology (Gosling & Mason, 2015) and market research (Comley, 2015). Online data collection has several advantages over laboratory studies, including lower infrastructure cost, faster and cheaper data collection (Casler, Bickel, & Hackett, 2013), and more extensive distribution of the study (Kan & Drummey, 2018). Apart from the previously discussed issues regarding measures in HCI research, concerns have been raised that data in online studies is frequently of low quality. For instance, Maniaci and Rogge (2014) and Meade and Craig (2012) have demonstrated that participant inattention can be a problem. Participants can provide invalid data in several ways. For example, content-responsive faking, which means that participants either change their answers to provide a certain image (present themselves in a better light), or they can present symptoms worse than they actually are. Another example is participants sometimes providing answers that are not related to the content, including random responses, or patterned responses (such as selecting the middle category for all items). Although these are not new phenomena (such as lie scales in the MMPI-2, Berry et al., 1992), recent research has increased its focus on content-unrelated responding (Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012). One of the reasons for the increased interest in this phenomenon could be that with the advent of online data collection the distance between researchers and participants and anonymity have both increased,

which may support such behavior. When participants complete studies online in exchange for course credits or money, extrinsic motivation can result in participants minimizing the time spent on answering questions to maximize the reward. This problem is accentuated on crowdsourcing platforms such as Amazon's Mechanical Turk (MTurk) or FigureEight. On these platforms, a large population of participants (workers) is readily available for completing tasks in return for small remuneration. Crowdsourcing platforms were initially created for small tasks that were difficult for computers to solve (Behrend, Sharek, Meade, & Wiebe, 2011). For instance, identifying certain objects (such as a cat) in images is sometimes difficult for computers. To improve the computing performance, large sets of validated training data for machine learning algorithms are needed. In addition to their success in computer science, crowdsourcing platforms have quickly gained the interest of researchers trying to efficiently recruit large samples for their studies (Behrend et al., 2011). Many works on crowdsourcing for psychological studies were positive in tone, suggesting it is a viable (and more diverse) alternative to other convenience samples (e.g., Casler et al., 2013; Kan & Drummey, 2018; Landers & Behrend, 2015; Paolacci & Chandler, 2014). Although there is research on the quality of survey responses collected on MTurk (Gadiraju, Kawase, Dietze, & Demartini, 2015), little is known about the performance of common methods for detecting inattentive respondents (Curran, 2016).

Inattentive responding is often referred to as content nonresponsivity, or more commonly, careless responding (Meade & Craig, 2012). Careless responding can be defined as answers that are unrelated to the content of a given item (Meade & Craig, 2012). It is usually present in situations where participants want to complete the survey as quickly as possible. It is important to note that answers can be close to random, but also distinctively non-random, such as when the same answer is selected for each item (such as the mid-point for each item), or when items are selected to form a pattern (such as 1, 2, 3, 4, 5, 4, 3, 2, 1[...]) Recent estimates of carelessness in online surveys range between 3% and 12% (Maniaci & Rogge, 2014; Meade & Craig, 2012), depending on the detection method and participant recruitment platform. Even low levels of carelessness may lead to failed replications (Oppenheimer, Meyvis, & Davidenko, 2009), including false-positives (Huang, Liu, & Bowling, 2015), failed experimental manipulations (Maniaci & Rogge, 2014), or problems with scale properties (D. Johnson et al., 2018; Kam & Meyer, 2015). Despite recent research efforts, estimates of carelessness in crowdsourced samples remain largely unknown. Most studies investigated mixed online samples (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012) or assessed only one type of carelessness measure (Dogan, 2018).

Hence, many questions concerning crowdsourcing and carelessness are still unanswered. For instance, little is known about the task-dependence and stability of carelessness. If participants respond carelessly in a survey, do they also answer carelessly in other tasks?

Further, frequency of carelessness in crowdsourcing tasks on various platforms is unknown, because most of the research focused on university participant pools and MTurk. However, until recently MTurk was only available for US residents. In contrast to MTurk, FigureEight allows researchers from various locations to distribute their surveys on several crowdworking platforms, without having to address them individually. However, the workers recruited on FigureEight might be more prone to carelessness, because the platform offers fewer community management features compared to Amazon (such as reputation management tools). Another issue is that carelessness cannot be determined with absolute certainty, and it remains debatable which method (or combination of methods) is most appropriate for filtering out such responses. Therefore, Curran (2016) proposed several new and more general measures, such as Person-total correlation or Resampled individual reliability. However, these still need to be examined empirically. Accordingly, the detection of carelessness with various methods, effects of excluding careless participants, and practical recommendations are the topics of manuscript 4 (Brühlmann, Petralito, Aeschbach, & Opwis, 2018).

Summary of the Manuscripts

The following manuscripts constitute this thesis. The first and second manuscripts have already been published, whereas manuscripts 3 and 4 are under review.

1. **Brühlmann, F.**, Vollenwyder, B., Opwis, K., & Mekler, E. D. (2018). Measuring the “why” of interaction: Development and validation of the user motivation inventory (UMI). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. (pp. 106:1–106:13). New York, NY, USA: ACM.
doi: 10.1145/3173574.3173680
2. Bargas-Avila, J. A. & **Brühlmann, F.** (2016). Measuring user rated language quality: Development and validation of the user interface language quality survey (LQS). *International Journal of Human-Computer Studies*, 86, 1-10.
doi: 10.1016/j.ijhcs.2015.08.010
3. **Brühlmann, F.**, Petralito, S., Rieser, D. C., Aeschbach, L. F., & Opwis, K. (2018). *TrustDiff: Development and validation of a semantic differential for user trust on the web*. Manuscript submitted for publication.
4. **Brühlmann, F.**, Petralito, S., Aeschbach, L. F., & Opwis, K. (2018). *Half of the participants in online surveys respond carelessly: An investigation of data quality in crowdsourced samples*. 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 of them will be referenced in the introduction and general discussion sections.

- Pimmer, C., **Brühlmann, F.**, Odetola, T. D., Oluwasola, D. O., Dipeolu, O., & Ajuwon, A. J. (2019). Facilitating professional mobile learning communities with instant messaging. *Computers & Education*, *128*, 102-112. doi: 10.1016/j.compedu.2018.09.005
- **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: Oxford University Press. doi: 10.1093/oso/9780198794844.003.0009
- Buttrick, N., Aczel, B., Aeschbach, L. F., Bakos, B. E., **Brühlmann, F.**, Claypool, H., ... Wood, M. (2018). Many Labs 5: Registered replication report of Vohs and Schooler (2008), Study 1. Manuscript submitted for publication.
- Ebersole, C. R., Chartier, C. R., Hartshorne, J. K., IJzerman, H., Mathur, M. B., Ropovik, H., ... **Brühlmann, F.**, ... Nosek, B. A. (2018). *Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability*. Manuscript in preparation.
- 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. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. New York, NY, USA: ACM. doi: 10.31234/osf.io/u94qt
- Pimmer, C., **Brühlmann, F.**, Odetola, T. D., Dipeolu, O., Gröhbiel, U., & Ajuwon, A. J. (2018). Instant messaging and nursing students' clinical learning experience. *Nurse Education Today*, *64*, 119–124. doi: 10.1016/j.nedt.2018.01.034
- Vollenwyder, B., Iten G. H., **Brühlmann, F.**, Opwis, K., & Mekler, E. D. (2018). *Salient beliefs influencing the intention to consider web accessibility*. Manuscript submitted for publication.
- Vollenwyder, B., Schneider, A., Krueger, E., **Brühlmann, F.**, Opwis, K., & Mekler, E. D. (2018). How to use plain and easy-to-read language for a positive user experience on websites. In *Proceedings of the 16th International Conference on Computers Helping People with Special Needs*. (pp. 514–522). Linz, Austria. Wiesbaden: Springer.

- **Brühlmann, F.** (2017, March 23). Can we trust big five data from the WVS? [Blog]. <https://bruehlmann.io/blog/dataquality/2017/03/23/Can-we-trust-big-five-data/>.
- Petralito, S., **Brühlmann, F.**, Iten, G., Mekler, E. D., & Opwis, K. (2017). A good reason to die: How avatar death and high challenges enable positive experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. (pp. 5087–5097). New York, NY, USA: ACM. doi: 10.1145/3025453.3026047
- Mekler, E. D., **Brühlmann, F.**, Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, *71*, 525-534. doi: 10.1016/j.chb.2015.08.048
- **Brühlmann, F.**, Schmid, G.-M., & Mekler, E. D. (2016). Online playtesting with crowdsourcing: Advantages and challenges. *CHI 2016 Workshop: Lightweight Games User Research for Indies and Non-Profit Organizations*. Retrieved from http://gur.hcigames.com/wp-content/uploads/2016/05/CHIGUR2016_paper_6.pdf.

Manuscript 1: Measuring the "Why" of Interaction: Development and Validation of the User Motivation Inventory (UMI)

Motivation and aim of the study. Developing, reflecting about, and extending measures is important for theory-building (Bringmann & Eronen, 2016). Hence, to contribute to the better understanding of the effects of user motivation on their experience, a multidimensional scale measuring user motivation based on SDT was developed. Apart from the main goal of contributing to the understanding of motivation in the field of user experience research, proximal goals of this research were to report scale development transparently, and to create a template for future scale development endeavors in the community (Kay et al., 2017). To fulfil the aim of transparency, all material (instructions and survey), analysis scripts, and data sets of both studies have been made available online¹. Additionally, the resulting paper was published under an open access creative commons license.

Development and validation strategy. The development and validation of the questionnaire followed best practice (DeVellis, 2016; Moosbrugger & Kelava, 2007) and consisted of four different phases. First, a review of existing scales of motivational regulations based on SDT from various domains of application was conducted. Items were extracted and rephrased from a diverse set of existing questionnaires in the areas of academic achievement (SIMS, Guay et al., 2000), video games (GAMS, Lafrenière, Verner-Filion, & Vallerand, 2012), sports (BRSQ, Lonsdale, Hodge, & Rose, 2008; BREQ Mullan, Markland, & Ingledew, 1997 and BREQ-2, Markland & Tobin, 2004; SMS-6, Mallett, Kawabata, Newcombe, Otero-Forero, & Jackson, 2007; SMS-II, Pelletier, Rocchi, Vallerand, Deci, & Ryan, 2013; PLOC-R, Vlachopoulos, Katartzis, Kontou, Moustaka, & Goudas, 2011), environmental protection (METS, Pelletier, Tuson, Green-Demers, Noels, & Beaton, 1998), romantic relationships (CMQ, Blais, Sabourin, Boucher, & Vallerand, 1990), therapy motivation (CMOTS, Pelletier et al., 1997), school (PLOC, Ryan & Connell, 1989), and well-being (Sheldon, Ryan, Deci, & Kasser, 2004). This item pool was reduced and refined. Second, the items were tested with a development sample in study 1. The goal was to optimize scale length and identify the best items for each of the six motivational regulations. In the third phase, the dimensionality, reliability, convergence, and discriminant validity were examined in an independent validation study. Finally, criterion validity of the UMI was investigated with participants who had thought about abandoning a technology.

Method Study 1. An item pool of 150 items was created, reviewed in an item sort task (Howard & Melloy, 2016) by the authors, then examined and refined by two psychologists with expertise in SDT who were not related to the study. The aim of the first step was to create an over-representative pool of items, and then further reduce them while assuring content validity through an expert review. This initial set consisted of 93 items. These

¹<https://www.usermotivation.org>

items were then tested with a development sample in study 1 to optimize scale length and to identify a subset of the best items for each of the six motivational regulations. Participants were recruited on Amazon Mechanical Turk ($N = 507$) and asked to name an interactive technology that they used frequently. Next, they answered a few questions relating to this technology and several scales (including the UMI). Data was then cleaned based on four measures: wrong answer to an instructed response item, less than four minutes to complete the survey, suspiciously large portion of items answered with the same value, and a negative Person-total correlation (see manuscript 4 for more details on this measure). A total of 481 participants, 39.1% male, with a mean age of 38.31 years ($SD = 12.61$) were included in the analysis. A majority of 33% chose to report their motivation for using Facebook. Other mentioned technologies included various smartphones, fitness trackers, handheld devices, or video game consoles. Only 15% reported that they used the technology once per day or less frequently.

Results Study 1. Item analysis with data collected in study 1 indicated that one item displayed unsatisfactory variance (less than 1). Two additional items were removed because their discriminatory power was below the recommended value of .30 (Borg & Groenen, 2005). For each construct, inter-item correlations and homogeneity were investigated. Six items were subsequently removed because their homogeneity was below .4. For the remaining 83 items, an exploratory factor analysis with principal axis factoring and oblimin rotation was conducted. In line with OIT, the number of factors to retain was set to six. Based on the results of the first exploratory factor analysis, communalities, primary- and cross-loadings were investigated to remove items with subpar properties (DeVellis, 2016; Howard, 2016). Results helped to reduce the number of items to 18 best-fitting candidates. A second exploratory factor analysis indicated that these items measure six distinct but related dimensions that follow the structure proposed by OIT: Conceptually close regulations correlate more strongly than conceptually distant regulations. The resulting scale and its measurement model was then tested in study 2.

Method Study 2. In study 2, the 18 mentioned items were tested with an independent sample of 460 participants. As with study 1, participants could complete the questionnaire based on any technology they had used frequently in the last 14 days. Apart from several questions related to the technology, the UMI and a selection of related UX and SDT scales were applied. Need satisfaction of autonomy, competence, and relatedness, which are core constructs of SDT (Ryan & Deci, 2000) were assessed using three items for each need, slightly adapted from Sheldon, Elliot, Kim, and Kasser (2001). Vitality after technology use, an important proximal measure of wellbeing, was measured using seven items of the state vitality scale by Ryan and Frederick (1997). As a more distal measure of wellbeing, life satisfaction was measured with the five items developed by Diener et al. (1985). In

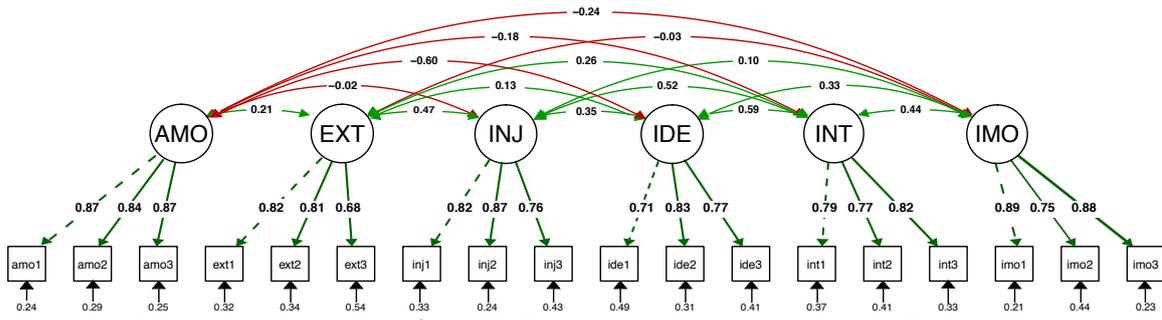


Figure 2. Measurement model of the UMI with standardized loadings. Dotted lines indicate loadings that were constrained to one. [$\chi^2_{120} = 237.53$, $p < .001$, $\chi^2/df = 1.98$, $CFI = .966$, $SRMR = .046$, $RMSEA = .046$, $PCLOSE = .771$]

terms of UX related measures, Usability was measured using the Usability Measure for User Experience (UMUX) (Finstad, 2010) and Likelihood to recommend was measured with the single item commonly used to calculate the Net Promoter Score (Reichheld, 2003; Sauro & Lewis, 2012).

Results Study 2. Confirmatory Factor Analysis (CFA) was conducted to test the proposed multidimensional factor structure of the UMI. Because multivariate normality was not given, robust maximum likelihood estimation with Huber-White standard errors and a Yuan-Bentler based scales test statistic were used. Results suggested that the proposed model fits the data well ($\chi^2_{120} = 237.53$, $p < .001$, $\chi^2/df = 1.98$, $CFI = .966$, $SRMR = .046$, $RMSEA = .046$, $PCLOSE = .771$). Standardized loadings and covariances are depicted in Figure 2. A model comparison revealed that a two or three factor model did not outperform the fit of the six factor model.

Reliability, convergent, and discriminant reliability of the subscales was investigated with congeneric reliability and internal consistency (Cronbach's alpha). The Average Variance Explained (AVE) was above the threshold of 0.5, suggesting high convergent validity, and the maximum shared variances were lower for each subscale than the corresponding AVE scores, which indicates discriminant validity. The relationship of the UMI and its six dimensions with other scales was investigated. The general pattern indicated that more self-determined regulations correlated more strongly with need satisfaction and vitality. Life satisfaction was not correlated with the UMI, which is not surprising given that the motivational regulation of a single, frequently used technology is distal of a more general satisfaction with life. In terms of UX measures, usability was negatively associated with amotivation and external motivation, and positively associated with identified and intrinsic motivation. Neither introjection nor integrated regulation were significantly correlated with usability. Likelihood to recommend was positively associated with the more self-determined regulations, and negatively related to amotivation. However, it was not correlated with ex-

ternal regulation. As an initial test of criterion validity, we investigated whether the UMI was able to detect differences between those participants who had questioned their use at some point, and those who had never questioned their technology use. Participants were divided in two groups based on their answer to question 4 (“Have you ever thought about quitting using [this technology]?”). We found that 163 participants questioned their use of technology at some point. Yuen-Welch tests on all six subscales of the UMI demonstrated that except for introjected and external regulation, all differences in motivational regulations were statistically significant with small, moderate, and large effect sizes (see Table 1).

	Use never questioned ($n = 297$)			Use questioned ($n = 163$)			Yuen-Welch test			
	M	SD	M_{tr}	M	SD	M_{tr}	t	df	p	ξ
AMO	1.74	1.054	1.38	2.98	1.649	2.79	-8.496	122.9	< .001	0.583
EXT	2.05	1.353	1.62	2.23	1.381	1.89	-1.831	178.6	.069	0.124
INJ	2.46	1.670	2.00	2.18	1.375	1.82	1.081	244.7	.281	0.085
IDE	5.38	1.342	5.54	4.58	1.473	4.58	5.849	180.2	< .001	0.404
INT	3.69	1.665	3.67	3.15	1.512	3.05	3.567	223.0	< .001	0.250
IMO	5.90	1.176	6.16	5.65	1.030	5.72	3.526	208.8	.001	0.257

Table 1

Comparison of participants who never questioned their use and participants who thought about quitting. M_{tr} = 20% trimmed means used for the Yuen-Welch test. ξ = Explanatory measure of effect size; interpretation: 0.10 small, 0.30 medium, 0.50 large.

Discussion and conclusion. The present work describes how a measure of user motivation was developed and validated. The UMI is rooted in SDT, and it was developed with a mixture of top-down and bottom-up approaches (meaning reuse items of existing scales). Results from both studies suggest that the UMI follows the proposed factor structure and measures six different motivational regulations reliably and validly. Correlations of the UMI with related measures follow existing SDT research, most notably on the relationship between need satisfaction and motivation.

Users indicated relatively high levels of the more self-determined motivational regulations (such as identified, integrated, and intrinsic motivation), which may reflect the leisure-oriented technologies participants decided to report on. During spare time, technology use may be much more driven by interest and enjoyment, and accompanied by a feeling of autonomy, compared to other contexts (such as at work). Approximately one third of the participants indicated they had thought about stopping using a technology. This group can be characterized by lower levels of intrinsic motivation, integrated and identified regulation, and higher levels of amotivation. Although thinking about quitting may not directly lead to actually abandoning a technology, research on the motivation of high school students

indicated that less self-determined motivation correlated with higher levels of drop-outs one year later (Vallerand, 1997). Thus, the UMI may help to identify users that are at risk of abandoning a product. The UMI may also help to understand if and how technology affects user well-being, because higher levels of self-determined regulations are associated with higher vitality. Additionally, autonomy supportive design (Calvo, Peters, Johnson, & Rogers, 2014) can be evaluated with the UMI to understand how it may influence motivation more successfully. A central limitation of the UMI is that it was developed to measure “technology use” in general, and not tied to a specific experience. The reason for this was that the wording of many existing scales is connected to specific life domains or activities rather than single episodes. Further, UX research emphasizes the importance of studying single experience episodes (Hassenzahl & Tractinsky, 2006). Therefore, in a next step, the UMI should be adapted to measure motivation on an experience level. Results from other domains with scales for situational motivational regulation are encouraging (e.g., Guay et al., 2000).

The UMI fills an important research gap, as it measures motivational regulations based on OIT, a subtheory of SDT. It is grounded in theory; therefore, existing evidence and theoretical models may be applied to study the strengths and weaknesses of SDT in the context of UX. The items of the UMI are deliberately general; to ensure the measure applies to various settings and products. While further research is needed to establish the UMI as a validated measure, the reported psychometric properties are encouraging.

Manuscript 2: Measuring user rated language quality: Development and Validation of the user interface Language Quality Survey (LQS)

Motivation and aim of the study. Reviews by expert translators or linguists are often regarded as the best way of ensuring consistent high quality. However, such reviews are expensive and time-consuming, especially when a global market is targeted. For instance, in 2016, the YouTube user interface was available in 60 languages, often rendering reviewing all changes for all languages by experts impossible. Therefore, a user-focused evaluation method was needed to identify the most urgent problems, and to appoint experts efficiently. In this practice-oriented context, a bottom-up scale development strategy is appropriate because specific requirements can be taken into account, and generalizability and theory building is less of a focus. However, it is crucial to create a valid and reliable tool that can be used in various languages and with several different products.

Item development. The development of the initial item pool followed a bottom-up approach, because there was no accepted theory or model of language quality in user interfaces. A group of linguists assembled in a brainstorming session and developed a set of criteria for good interface language quality. The items of the questionnaire were then derived by the first author from the following criteria: friendliness, casualness, professionalism, naturalness, ease-of-understanding, appropriateness, correctness, and global satisfaction.

Method Study 1. The goal of study 1 was to administer the scale to a test sample and identify the strengths and weaknesses of the items. English-speaking users on the YouTube platform were invited to participate in the study. Users were asked to rate the text quality of the YouTube interface, with all 10 items presented in sequential order. The sample ($N = 3588$) was subjected to a rigorous cleaning procedure to make sure that participants actually rated user interface text, were native English-speakers, frequently interacted with YouTube, and used its interface in English.

Results Study 1. After data cleaning, 843 responses remained and were included in the analysis. The majority were male (73.5%), and 55.4% were between 18 and 29 years old. Participants tended to answer the items with the upper part of the scale, showing left-skewed distributions. Discriminatory power of each item, and the corresponding homogeneity, was satisfactory except for item 2 “How casual or formal is the text used in the [product name] interface?”. With the open-ended questions at the end of the questionnaire, we learned that this item is difficult to interpret because casualness and formality are highly subjective aspects and might be perceived and judged very differently by different users. Therefore, it was decided to remove item 2. The qualitative data also suggested that users relatively frequently encountered text that did not make sense (in their opinion). Hence, a new item that would allow measuring the occurrences of nonsensical text was included: “How often do you encounter text that does not make sense?”.

Method Study 2. In the second study, a revised version of the 10 item scale was applied ($N = 3327$). As with study 1, the same data cleaning procedure was used. Accordingly, 2161 participants were excluded because they indicated that they rated the language quality of user-generated content. In the next step, 333 participants reported that English was not their native language, 7 did not use YouTube at least once a week, 95 used YouTube with other languages, and 41 participants either left more than half of the items unanswered or answered all questions with the same value. The final data set included 690 respondents.

Results Study 2. In study 2, which included the 690 participants, results of the item analysis indicated sufficient discriminatory coefficients and homogeneity indices. Internal consistency, as measured with Cronbach's α , was high at .820. An exploratory factor analysis with oblimin rotation was conducted to investigate the structure of the scale. Based on the Kaiser criterion (eigenvalue > 1) two factors were identified that explained 58.2% of the variance. The emerged factors correlated with $r = .429$. An interpretation of the factor loadings suggested that the first factor described *Linguistic Correctness* and the second factor described *Readability*.

Results additional studies. In a third study, validity and generalization were examined through an investigation of correlations of the LQS with UMUX (Finstad, 2010). Correlations of the overall LQS score with the UMUX was moderate ($r = .396$, $p < .01$, $N = 211$) and the Readability subscale correlated stronger ($r = .446$, $p < .01$, $N = 211$) than the Linguistic Correctness subscale ($r = .157$, $p < .05$, $N = 211$). Discriminative validity was examined by comparing LQS scores of participants who rated user-generated content to participants who rated the interface text. Results indicated that they rated the language quality significantly lower than those who rated the user interface text, $t(752.184) = 15.645$, $p < .001$, $d = 0.99$. In the next step, the LQS was translated into nine languages and its item statistics and psychometric properties were investigated for each of ten different regions. Difficulty indices, discriminatory power, homogeneity, and internal consistency were in a similar range as in study 2. The LQS was also applied to Google Analytics and AdWords, achieving satisfactory results. This indicated that the LQS can be applied in various languages and for different products.

Discussion and conclusion. This paper presents the development and validation of the LQS, a reader-focused evaluation method for user interface text. It allows companies to source their users to rate the language quality in a user interface, subsequently increasing efficiency with expert evaluations and reworking of user interface text. With two studies, the final version of the scale was developed and refined. The final LQS displayed good psychometric properties, and an exploratory factor analysis demonstrated that the LQS measures two distinct but related facets *Linguistic Correctness* and *Readability*. Content validity was assured by involving expert linguists in the process, and criterion-related validity was

measured using correlations with the global item. Convergent validity was demonstrated through moderate correlations of the LQS with usability, and discriminative validity was investigated with a comparison between participants who rated user-generated content and participants who rated the expert-created user interface text. Results from studies in languages such as Spanish, German, or Arabic, as well as studies with other Google products, exhibited promising psychometric properties. According to del Galdo and Nielsen (1996) there are three levels for approaching the problem of international user interfaces. The first level is the correct technical implementation of the user native language character set, including notations and formats. The second level is designing a user interface and user information that are understandable and functional in their native language. At this level, the LQS can help practitioners receive user feedback about linguistic correctness and readability, which helps to prioritize resources and improve user experiences with an interface. This is the basis for the third level of internationalization: Designing interfaces that address specific cultural models, such as the way people communicate or the way business is conducted in different cultures. The LQS can be applied at various stages of design and development to track and improve user experiences with an interface language.

The presented studies are also subject to limitations. First, the validation of the LQS is not finished, as it requires more independent investigations in other domains to identify specific limits and strengths. Further, the LQS has only been developed and validated with websites on desktop computers, and needs to be tested with mobile applications to ensure broad applicability. Second, future research should also include more objective measures (such as error rates or expert judgment), and then correlated with LQS scores to further study its validity. Lastly, participation in the reported studies was “opt-in.” Therefore, the sample is self-selected and might include a sampling bias. This issue is important for the interpretation of the results, because they might not reflect a representative perception of users.

Manuscript 3: TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web

Motivation and aim of the study. The goal of this project was to develop a short and versatile measure of user trust in English with good psychometric properties. As a first step, existing questionnaires following the models of benevolence, competence, and integrity, were reviewed and items were collected (Bart, Shankar, Sultan, & Urban, 2005; Bhattacharjee, 2002; Cho, 2006; Corbitt, Thanasankit, & Yi, 2003; Flavián et al., 2006; Gefen, 2002; Gefen, Karahanna, & Straub, 2003; Hong & Cho, 2011; Jian, Bisantz, & Drury, 2000; Kofaris & Hampton-Sosa, 2004; Lu et al., 2012; McCroskey & Teven, 1999; Pavlou & Gefen, 2004; Rieser & Bernhard, 2016). Because most of these items used adjectives to describe certain aspects (such as “I think that the information offered by this site is *sincere* and *honest*”), these words were then extracted and sorted according to the overarching construct. Subsequently, for each of the 43 unique adjectives, several antonyms were selected with the help of dictionaries (www.merriam-webster.com, www.thesaurus.com, www.leo.org). After this, 28 positive adjectives with up to 3 antonyms remained. With this initial set of items, a review was conducted using 18 trained psychologists and HCI researchers. In an online survey, experts assigned each word to one of the three dimensions of trust: benevolence (BEN), integrity (INT), and competence (COM). The critical value for an item sort task with 18 experts was 13, thus items that were correctly assigned by less than 13 experts were excluded (Howard & Melloy, 2016).

Three studies were conducted to validate the questionnaire, and each study served different purposes: Study 1 reduced the item pool and identified the best candidate items, study 2 tested the measurement model of this questionnaire in a different setting, and study 3 conducted an initial test of criterion validity with an experiment.

Method Study 1. The goal of study 1 was to reduce the over-representative item pool by employing exploratory factor analysis, and to test the convergent and discriminant validity of the scale. A total of 714 participants successfully completed the online survey on Amazon’s Mechanical Turk platform. Participants were excluded if the response time was under 150 seconds, if a response pattern such as repeated selection of the same values was present, or if participants indicated that we should not use their data at the end of the survey. After this procedure, data from 601 participants remained (42% women, mean age = 38 years, age range 18–84). In the study, participants were asked to complete two tasks on one of two randomly assigned websites. Both websites were in the English language and relatively unknown in the US. When participants returned to the survey, they were asked to fill in 20 items of the TrustDiff, Likert-type Trust scale (Flavián et al., 2006), visual aesthetics of websites inventory (Moshagen & Thielsch, 2010), and the UMUX (Finstad, 2010).

Results Study 1. The main goal was an item reduction process. First, the distribution statistics for each item were examined, indicating that three items were slightly negatively skewed. All three items were part of the competence factor, which was measured with nine items. These items were then excluded to balance the three subscales. Exploratory factor analysis with oblique rotation fixed to extract 3 factors was conducted with the remaining 17 items. Three items had to be excluded on the grounds of high cross-loadings or insufficient loadings on the designated primary factor. In a second exploratory factor analysis, the remaining 14 items were included and displayed high primary loadings and low cross-loadings. These 3 factors explained 74% of the variance, and the internal consistency of each scale was significantly above the threshold of .70. Correlations of the TrustDiff subscales with Trust measures with the items from Flavián et al. (2006) were high, with each subscale correlating most strongly with the other subscale. Usability and visual aesthetics correlated moderately with the 3 subscales of the TrustDiff (.33–.53), which was slightly lower than the Likert-type trust scale. This refined questionnaire was then tested with a confirmatory factor analysis and a different setup in study 2.

Method Study 2. The goal of this study was to test the proposed factor structure of the revised TrustDiff. Participants were asked to name a single interactive technology that they use frequently. The remainder of the study focused on this particular technology, and the 14 items of the TrustDiff were included. A total of 315 participants completed the relevant part of the study. Three participants had to be excluded because they indicated that their data should not be used, resulting in a final sample of 312 participants (44% men, mean age = 37.6 years, age range 18–76). The most frequently chosen technology was Facebook (42.7%), followed by other types of social media, Fitbit, and Microsoft Word or Excel.

Results Study 2. To test the three-dimensional factor structure, a confirmatory factor analysis was conducted. Multivariate normality was not given, therefore a robust maximum likelihood method with Huber-White standard errors and a Yuan-Bentler based scaled test statistic was used. Results with all 14 items resulted in an acceptable fit, $\chi^2(74) = 140.530$, $p < .001$, $\chi^2/df = 1.89$, $CFI = .971$, $SRMR = .047$, $RMSEA = .054$, $PCLOSE = .279$. Modification indices proposed additional covariance between certain items of a subscale. However, because the goal was to have a parsimonious scale, removing items was preferred. Hence, four items were excluded based on statistical and theoretical grounds. The resulting scale with 10 items measured 3 related but distinct dimensions, and displayed excellent psychometric properties, $\chi^2(32) = 32.500$, $p = .442$, $\chi^2/df = 1.02$, $CFI = 1.000$, $SRMR = .027$, $RMSEA = .007$, $PCLOSE = .996$ (see Figure 3). Study 2 demonstrated that the scale could be reduced without losing reliability.

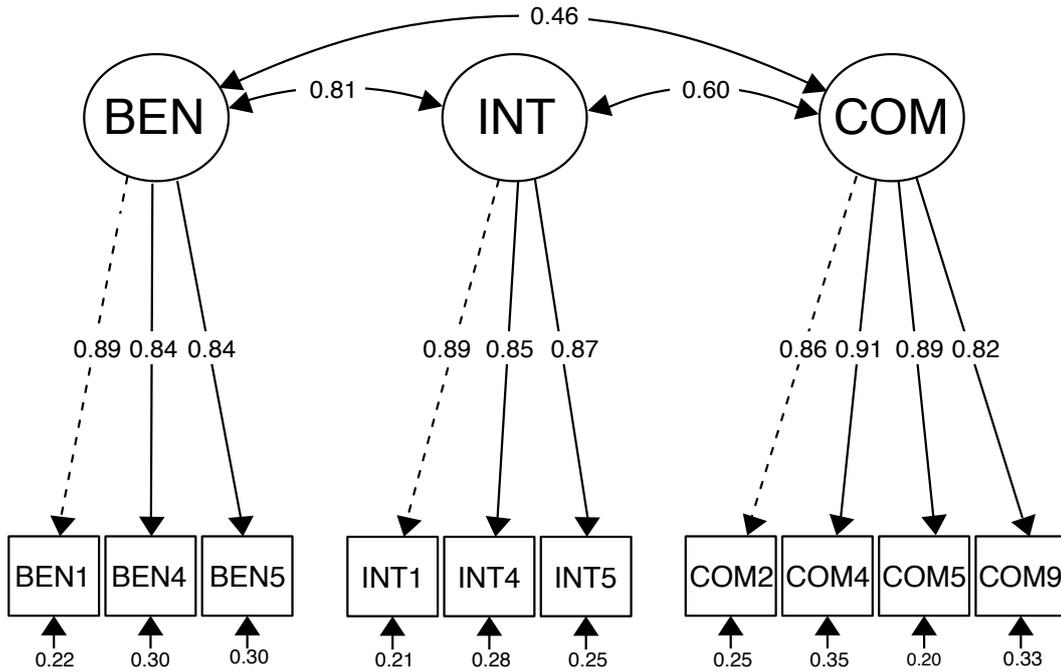


Figure 3. Measurement model of the TrustDiff with standardized loadings [$\chi^2(32) = 32.500$, $p = .442$, $\chi^2/df = 1.02$, $CFI = 1.000$, $SRMR = .027$, $RMSEA = .007$, $PCLOSE = .996$].

Method Study 3. After development study 1 and validation study 2, the third study was concerned with criterion-related validity. As part of a larger study, participants were asked to rate a mock online shop with the TrustDiff, based on a screenshot. Participants were randomly assigned to one of two versions of the mock online shop. The first group was shown a screenshot of an online shop that included several trust-supporting elements (high trust) while the second group was given a screenshot of the same shop without these elements (neutral). The elements were manipulated according to the trust related elements described in Seckler et al. (2015). Participants had to examine the online shop for at least four seconds to continue in the study. After that, they had to complete in the TrustDiff, Likert-type scale of trust (Flavián et al., 2006), and visual aesthetics (VisAWI, Moshagen & Thielsch, 2010) of the website. A total of 394 participants from the US were recruited on FigureEight, who subsequently completed the relevant part of the survey. Data was cleaned with two attention check items, which reduced the sample size to 258. Six additional participants were excluded because they indicated that we should not use their data. The final sample included 252 participants (28% men, mean age = 39 years, range 18–78).

Results Study 3. On average, participants viewed the websites for 1.47 minutes. No significant differences in viewing time were observed between the conditions, $t(246.88) =$

Table 2

Descriptive statistics and results of Welch's two samples t-test as an assessment of criterion validity of the TrustDiff.

		High trust (<i>n</i> = 128)		Neutral (<i>n</i> = 124)		t	df	<i>p</i>	d
		M	SD	M	SD				
TrustDiff	Benevolence	5.01	0.962	4.28	1.075	5.681	245.0	< .001	0.72
	Integrity	5.48	0.993	4.75	1.199	5.210	238.7	< .001	0.66
	Competence	5.58	1.011	4.47	1.455	6.989	218.6	< .001	0.89
	Total	5.37	0.899	4.50	1.176	6.577	230.1	< .001	0.84

Note. Total *N* = 252.

0.073065, $p = .9418$. All measures deviated significantly from a normal distribution; therefore, both Welch's two samples t-test and robust Wilcoxon rank-sum test were conducted. Both tests resulted in the same conclusions for all measures; therefore, only the results of the Welch's t-test were reported. As presented in Table 2, statistically significant differences between the two conditions were observed for all subscales of the TrustDiff indicating criterion validity.

Discussion and conclusion. The aim of this study was to develop and validate a short scale to measure trust in websites. This study contributes to the measurement of trust with a broadly applicable semantic differential, which extends existing scales that are more tailored to specific sites. The word pairs of the TrustDiff merely comprise of two antonyms that should (ideally) be easily translated into languages other than English. The proposed 10-word pairs were constructed based on existing literature, and tested for linguistic and psychological bipolarity by an expert panel. Results from exploratory factor analysis in study 1 suggested a 14-item scale with three distinct but related trust dimensions. In study 2, these 14 items were again tested for structural validity in a different context. Results indicated that the scale could be reduced to 10 items without losing reliability and maintaining excellent psychometric properties. Moreover, study 3 demonstrated that the TrustDiff is sensitive to websites with differences in trust-related features. The three studies presented here entail an initial validation for the TrustDiff. The scale exhibited promising psychometric properties, but it needs to be further tested in-depth with various products and services in different contexts. The development and validation followed best practices, and the scale is readily applicable to varied research contexts.

The TrustDiff contributes to research on online trust with a scale that is applicable to a broad range of products. Compared to the existing context- and language- specific questionnaires (e.g., Flavián et al., 2006; Lu et al., 2012; McKnight, Choudhury, & Kacmar, 2002a) the TrustDiff can be applied in various contexts of user trust on the web. The item pairs may be easier to translate into different languages than the relatively long statements

of Likert-type scales, and the semantic differential scale also offers a broader range of possible answers between the two semantic poles (Chin et al., 2008). The TrustDiff can assist further investigations into how web design elements relate to trust, but may also permit comparisons between different products. For instance, trusting beliefs of Facebook users can be compared to users of eBay by using TrustDiff. Apart from this practical contribution, results of the three-factor model in Figure 3 demonstrate that the latent variables benevolence and integrity correlated substantially in this context (.81). This overlap was also observed in study 1 of the fourth manuscript, albeit slightly less strongly, where the means of the benevolence and integrity subscales correlated with $r = .74$. This may indicate that these two constructs are not easily separable. However, the correlation between the benevolence and integrity subscales of the Likert-type scale (Flavián et al., 2006) was even stronger ($r = .84$); thus, it may be a problem of the three-factor model of trust. Hence, future research needs to investigate this notion and test alternative models of trust. Nevertheless, the TrustDiff offers several advantages over existing Likert-type trust scales, such as a broader and simpler application in different contexts and various products. Further, it may be simpler to translate into different languages. Additionally, it allows users to rate trust more fully from a negative to a positive pole in one, short, economical scale. With state-of-the-art development, and validation with over 1000 participants in three independent studies, the TrustDiff is a viable and readily applicable alternative to existing Likert-type questionnaires for determining user trust.

Manuscript 4: Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples

Motivation and aim of the study. In the present study, six different carelessness detection methods were assessed to study three research questions. Research Question 1) How prevalent is careless responding in samples from crowdsourcing platforms, based on various detection methods for carelessness? While it is challenging estimate carelessness, we followed the procedure of Meade and Craig (2012) to determine the number of careless participants through latent profile analysis. Research Question 2) How are task-specific measures of carelessness (such as open-ended questions) related to planned detection methods and post-hoc methods? We aimed to test how different detection methods overlap with open answer quality, an easy to implement and widely used measure in surveys that is not based on Likert-type scales. Research Question 3) Based on our findings, which methods are most applicable for identifying carelessness in a crowdsourced sample? To answer this question, six different carelessness measures were used to detect subgroups of participants with similar characteristics. Subsequently, a predictive model was built to efficiently identify the careless participant group.

Method. To detect carelessness, several special items were included in a survey about negative online shopping experiences. The so-called **planned detection methods** included aggregated answers from the self-reported items, a Bogus item, an Instructed response item (IRI), and response time (see Table 3). **Post-hoc detection methods** were LongString analysis, Odd-even consistency, Resampled individual reliability, and Person-total correlation (Curran, 2016). These measures do not need a special scale or item; rather, they can be calculated on any reasonably long scale. The study included an open question with a text area for participants to complete. The free-text answer given by the participants was rated in terms of quality, and was incorporated as a task-related measure of responding quality.

The study was set up as follows: After providing consent, participants recalled a recent negative experience with an online store and responded to two open-ended questions in a large text area. They were asked 1) to describe what caused this experience to be negative, and 2) how this affected their online shopping habits. Next, several Likert-type scales were presented as distractors. In the second part, participants were randomly presented one of two versions of a mock online shop. The online shop in both conditions was the same (a clothing store), but in the high trust condition it was enhanced with trust related features such as high quality images or trust seals, whereas these features were absent in the low trust condition. The aim was to conduct a plausible experiment that was thematically related to the rest of the study. After participants examined the shop, they completed 16 items of a Likert-type scale for trust in web vendors (Flavián et al., 2006). Later in the analysis, this

experiment was used to investigate the effects of excluding careless participants on effect sizes and p-values. On the next page, participants first responded to a visual aesthetics measure for the present website (VisAWI, Moshagen & Thielsch, 2010) and then the Big Five Inventory (BFI; O. P. John & Srivastava, 1999). All post-hoc detection methods of carelessness were applied to the 44 items of the BFI in the last part of the questionnaire. We decided to focus on the BFI because it is multidimensional, with sufficient length to calculate various indices, and it has also been the basis in several other studies in this field (J. A. Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012). On the last pages, participants filled in ten items based on Maniaci and Rogge (2014) to assess general tendencies in responding. Although excluding participants based on self-reported responding tendencies has been found to improve data quality significantly (Aust, Diedenhofen, Ullrich, & Musch, 2013), these items are also easily detected and prone to manipulation and dishonest answers. Three items were used to measure self-reported careless responding ($\alpha = .84$), two items to measure self-reported patterned responding ($\alpha = .88$), three items to assess self-reported rushed responding ($\alpha = .83$), and two items assessing self-reported skipping of instructions ($\alpha = .68$). Additionally, self-indicated data usage was assessed using the Self-reported single item (SRSI UseMe) developed by Meade and Craig (2012). Finally, regardless of the quality of their answers, all participants received a completion code.

Four-hundred participants were recruited on the crowdsourcing platform FigureEight, of which 394 provided complete data sets that were included in the analysis. Although participants were specifically recruited from the US, 12 participants indicated that they lived in other countries.

Results. The analysis was structured as follows: First, the different detection methods with their recommended cut-offs were investigated, both individually and in combination. Second, the relationship between Answer quality and other detection methods was investigated, along with the correlations between all methods. Third, the different methods were used to identify different classes of participants in the data to estimate the number of careless participants. Additionally, a predictive model was constructed to identify the most efficient methods for detecting this class.

The overview in Table 4 indicates that the majority of participants were flagged by at least one carelessness detection method (59.14%). Most participants were flagged by self-reported measures (26.90%), followed by Answer quality (25.38%).

Correlations indicated that overlap between the different methods was relatively low, and often lowest with Answer quality (.23–.26). This suggests that some of the methods may identify different types of carelessness, and that task-specific measures of carelessness (such as Answer quality) are important in ensuring that participants included in the analysis are attentive.

Table 3

The items of the self-reported responding tendencies scale (Maniaci & Rogge, 2014) and planned detection items included in the study. Self-report answer options ranged from 1 (never), over 4 (approximately half the time), to 7 (all the time). The Bogus item was included in the BFI, where answers between 1 (disagree strongly) and 5 (agree strongly) were possible. The IRI was included in the trust scale, which was used as the dependent variable of the experiment.

Measure	Item
Self-report	[How often do you...]
Careless responding	<ol style="list-style-type: none"> 1. Read each question 2. Pay attention to every question 3. Take as much time as you need to answer the questions honestly
Patterned responding	<ol style="list-style-type: none"> 4. Make patterns with the responses to a block of questions 5. Use the the same answer for a block of questions one the same topic [rather than reading each question]
Rushed responding	<ol style="list-style-type: none"> 6. Answer quickly without thinking 7. Answer impulsively without thinking 8. Rush through the survey
Skipping of instructions	<ol style="list-style-type: none"> 9. Skim the instructions quickly 10. Skip over parts of the instruction
SRSI UseMe	In your honest opinion, should we use your data in our analyses in this study?
Bogus item	[I see myself as someone who ...] Did not read this statement
IRI	I read instructions carefully. To show that you are reading these instructions, please leave this question blank.

As in (Meade & Craig, 2012), the raw values of all non-self-reported measures were included in a latent profile analysis to identify different classes of participants with similar answering patterns. Results revealed three classes of participants. Class 1 included 181 (45.9%) observations, and displayed a pattern that could be described as conspicuous. Compared to the remaining Classes 2 (129, 32.7%) and 3 (84, 21.3%), Class 1 exhibited lower Answer quality, lower self-reported quality, and more frequently failed the Bogus item and the IRI. Response time was also slightly lower, and Odd-even consistency, Resampled individual reliability, and Person-total correlation were lower than in the unsuspecting classes. Average LongString values were almost twice as high in Class 1 compared to Classes 2 and 3. The rate of carelessness in this crowdsourced sample (based on the latent profile analysis), was 45.9%. Hence, based on this figure, almost half of the participants could be described as inattentive. A conditional inference tree was built to predict class membership based on the different carelessness methods (Hothorn, Hornik, & Zeileis, 2006). The aim was to identify

Table 4

Descriptive statistics for all detection methods used in the study. Self-report includes problematic responding tendencies as well as the SRSI UseMe item.

	Mean	SD	Min	Max	No. Flagged	%
Planned detection						
Self-report					106	26.90
Bogus item					92	23.35
Instructed response item					96	24.37
Response time	16.71	9.22	3.93	61.15		
Post-hoc detection						
LongString	6.63	9.15	0	44	25	6.35
Odd-even consistency	.61	.43	-1	1	63	15.99
Resampled individual reliability	.56	.39	-.82	.99	63	15.99
Person-total correlation	.38	.32	-.47	.88	74	18.78
Answer quality					100	25.38
Total (flagged by at least one method)					233	59.14

Note. Total $N = 394$

a subset of methods that would successfully predict Class 1 membership. Results indicated that Class 1 could be efficiently predicted through Answer quality, the IRI, and the Bogus item. Hence, these three measures can be recommended. Additionally, we recommend the LongString index, because it was able to identify clearly suspicious patterns (same answer for every item), and offers an unambiguous interpretation.

Discussion and conclusion. Comparison to the study by Maniaci and Rogge (2014), the planned and post-hoc detection methods flagged a higher percentage of participants in the crowdsourced sample. In the present study, the IRI flagged 24.4% of participants and the Odd-even consistency was 16%, while in Maniaci and Rogge (2014) the rates for the same measures were 14% and 7%, respectively. The LongString index, however, flagged a comparable number of participants (6.3% against 6%) in both studies. In the relatively short study presented here, the rate of inattentiveness (as measured by the IRI and the Bogus item) was relatively high. Considering the recommendation of Meade and Craig (2012) to place one attention check item every 50 to 100 items, the number of conspicuous participants could even increase in longer studies. In the substantially longer study by Peer, Brandimarte, Samat, and Acquisti (2017), 73% of participants were flagged by at least one attention check. When all different detection methods were combined, the number of conspicuous participants was 59.14%. However, this combination might be prone to falsely identifying otherwise unsuspecting participants. Therefore, a latent profile analysis was conducted to identify classes of participants with similar patterns in the different carelessness detection methods. Results demonstrated that one class with 181 (45.9%) of participants could be described as careless. This class subsumes multiple forms of carelessness, and can

be characterized by low open answer quality, high rates of failing the IRI and Bogus item, and high levels of self-reported carelessness/responding tendencies. The LongString index was considerably higher for this class, while the Person-total correlation was low. This indicates that Class 1 is highly consistent, but achieves low congruence with the total sample. However, the consistency measures of Resampled individual reliability and Odd-even consistency were also associated with Class 3. Thus, these methods should be used with caution, as they might result in high false-positive rates. The estimate of 45.9% careless respondents is considerably higher than in similar approaches. For instance, Maniaci and Rogge (2014) and Meade and Craig (2012) identified approximately 2.2% to 11% as careless with their latent profile analyses. This indicates that online surveys with participants from crowdsourcing platforms might be more susceptible to these kinds of problems with data quality. Hence, it is vital to assess data quality and employ measures to reduce carelessness, and to report data cleaning transparently.

The Mathews correlations between open answer quality, self-reports, attention checks, LongString index, Odd-even consistency, Resampled-individual reliability, and Person-total correlation were at a medium to low level. Therefore, this supports the observation that carelessness could be task-dependent. In a study with a mixed sample, Maniaci and Rogge (2014) demonstrated that inattention in specific tasks of a survey, such as attentively watching a video, was not strongly related to standard carelessness detection methods.

Based on the results from the conditional inference tree, the IRI, Bogus item, and a task-specific measure (such as open answer quality) can be recommended for researchers and practitioners. In the present study, these methods were able to identify almost all participants of Class 1. Additionally, LongString analysis provided a different perspective, which might complement these methods. In cases where special items or tasks are difficult to implement (such as in surveys of very specific populations), a LongString analysis could be an essential reference point to clean data sets, because it offers a relatively objective interpretation and is not heavily dependent on sample properties. In addition to these methods, the SRSI UseMe item can also be recommended, because participants self-indicate their data as problematic and it allows participants to withdraw consent at the end of the study. Other post-hoc methods are not recommended, because they were not clearly associated with one class, and they were not predictive of Class 1 membership.

While this study offers an estimate of carelessness on crowdsourcing platforms, it may not be representative of all studies, and it might not readily transfer to other platforms (such as Amazon's Mechanical Turk). Depending on the platform, different community-management features might decrease the probability of carelessness. Therefore, a systematic analysis of different platforms with a standard procedure could overcome this limitation. The analysis conducted in the reported study followed standard procedures in research on

carelessness, but it is not the only way to detect problematic respondents. For instance, response time could be analyzed at an item level and integrated into a model; thus, helping to assess data quality while participants are completing the survey. In this way, the tool could autonomously recruit more participants to guarantee a certain predefined sample size of high-quality respondents. Furthermore, the exclusion of approximately half of the sample might have severe methodological and economic consequences. Hence, more research is needed on measures for reducing or preventing carelessness. The answer to the open questions was used to gain insight into the task-dependency of careless behavior in this study. Although such combinations of open questions and Likert-type scales are quite frequent, a more systematic analysis of different tasks and their relation to carelessness in surveys is needed. The recommendations in this study were largely based on the latent profile analysis, which included results from an open-question task. Future research will indicate whether this method achieves similar results with other tasks. Finally, the estimate of 45.9% carelessness in this sample is based on the careless Class 1 identified with the latent profile analysis. There are many other ways to determine the number of problematic respondents in a survey, and all these methods are (to a certain degree) approximate and uncertain.

General Discussion

The studies presented here describe the development and validation of three different questionnaires, and an investigation into data quality of online surveys. Each of the manuscripts extends existing knowledge and tools in their specific research area, and opened up new avenues for future research. They address the challenge of quantifying user experience with similar methods but different initial positions, and in different domains of HCI research. In the current section, general remarks and conclusions concerning the over-arching theme evaluation and development of subjective measures for UX research are discussed.

Validity and validation

With the focus on subjective experiences, UX research has introduced new challenges in measurement (Hassenzahl & Tractinsky, 2006; Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). However, measures in UX are frequently self-developed, ad-hoc, have unknown psychometric properties, and may be of questionable validity (Bargas-Avila & Hornbæk, 2011). Validity means that a measure actually measures what it purports to measure (Borsboom, Mellenbergh, & van Heerden, 2004). Showing that a scale is valid in this sense is very challenging, because this seemingly simple definition of validity has several consequences for questionnaire development and – more importantly – for the philosophical underpinnings of measurement (Borsboom et al., 2004). All three questionnaires included in this thesis were developed from the perspective of classical test theory and latent variable models. Although classical test theory has also been criticized for its inconsistency and unrealistic assumptions, it remains the current mainstream theory in test and questionnaire development (Borsboom, 2005; DeVellis, 2016). From a theoretical perspective, showing that a scale is valid needs a theory of how a construct (such as motivation) causes a response to a specific questionnaire item (Borsboom et al., 2004). Demonstrating this causal relationship is an issue that goes beyond the scope of this thesis, and challenges virtually all measures of psychology and HCI. The focus of this thesis is on practical considerations in development and *validation*; the process of showing validity. Based on the understanding of DeVellis (2016) and Moosbrugger and Kelava (2007), validity in an absolute sense cannot be achieved, but different methods may demonstrate partial validity of a scale. DeVellis (2016) distinguishes three types of validity which are elaborated below.

The most basic concept of validity is *content validity*. Content validity is given when the scale includes all relevant aspects of a phenomenon, and excludes aspects that belong to different constructs (DeVellis, 2016). This is typically studied with expert reviews of an item pool. All three questionnaires reported in this thesis included an expert review stage. This review stage separates all these questionnaires from ad-hoc questionnaires, which are usually scales that have not undergone independent review.

Frequently, at least one theoretically related and one theoretically unrelated questionnaire are included in validation studies. These correlations can be used to evaluate *construct validity*. For instance, in manuscript 2, the UMUX was correlated with the LQS to show that the LQS correlates moderately with the related concept of perceived usability. This aspect of validity (showing that constructs that are not supposed to be related are actually unrelated) (Moosbrugger & Kelava, 2007) is commonly complemented with additional correlations, such as in manuscript 1, where positive, negative or absence of relationships of the UMI with several other scales were investigated. Despite being criticized as 'circumstantial' validity (Borsboom et al., 2004), embedding a scale in such a nomological network makes sense, because it may indicate inconsistencies of the scale, and helps to situate it in existing measures and literature. In addition to correlations with existing measures, the structure of correlations among the items was investigated for all three scales. The items were subject to exploratory factor analysis, and the structure of the UMI and the TrustDiff were explicitly tested with confirmatory factor analysis. In publications with ad-hoc, self-developed scales, these statistical (or structural) validation methods are frequently missing (Bargas-Avila & Hornbæk, 2011).

For practitioners and researchers, it is often relevant to show how a measure relates to an important criterion, whether it predicts a behavior or experience in the future, or whether the scale is able to differentiate between known groups of participants, patients, or users (Moosbrugger & Kelava, 2007). DeVellis (2016) refers to this aspect of validity as *criterion-related validity*. In manuscript 1, we demonstrated that the UMI is able to differentiate between two groups of users. In manuscript 2, the scores on the LQS were statistically different between ratings of user-generated text and user interface text, and in manuscript 3, users provided significantly different ratings on the TrustDiff scale depending on their experimental condition. Such thorough investigations of psychometric properties and validity are rare in HCI (for a notable exception see Moshagen & Thielsch, 2010). Without validation attempts and (even more so) without reporting individual items (Bargas-Avila & Hornbæk, 2011), findings of studies with ad-hoc scales are of questionable value. Ad-hoc scales hinder generalization of findings and aggregation of knowledge. Further, without validation false conclusions about constructs that were not actually measured might be drawn. Ad-hoc scales are a consequence, but may also cause fragmented, atheoretical research (Liu et al., 2014) and a lack of replication (Hornbæk et al., 2014).

Theory and measurement

Validity applies to all types of measurements. Another important aspect of measurement in psychology is the distinction between atheoretical and theoretical measures (DeVellis, 2016). DeVellis describes theoretical measures as any measure with a theoretical

foundation in the sense that the response to an item is a reflection of the phenomenon being studied. Atheoretical measures (such as questions about sex or age) are directly measurable or observable qualities or quantities, whereas theoretical measures go beyond these direct responses in building a hypothetical construct.

In manuscripts 1 and 3, the items for the questionnaires were selected with a specific theoretical model in mind; thus, the participant's responses to these items are thought to be caused by a hypothetical construct. For example, it is assumed that intrinsic motivation causes the response to the item "I use [technology X] because it is enjoyable". Accepting the idea that a hypothetical construct is the cause of a specific behavior (response to an item) has consequences for scale validation. In such situations, the items of a scale are indicators of the underlying construct; therefore, it needs to be demonstrated that these items form a single dimension (DeVellis, 2016). In manuscript 3, the TrustDiff was designed as a scale with three different dimensions: benevolence, integrity, and competence. This structure was tested in study 2 of manuscript 3 with confirmatory factor analysis. Basing a scale on a model allows researchers to set constraints in the structure of correlations between items, which can then be tested in various situations; thus, quantifying the empirical evidence for the model's assumptions. In manuscript 1, the UMI covers the multifaceted construct of motivational regulations as posited by OIT, a subtheory of SDT. The scale was designed to measure six different (but related) motivational regulations. Confirmatory factor analysis demonstrated that this structure with six dimensions is a meaningful description of the data. Thus, the structural validity of both scales was tested and supported (DeVellis, 2016).

Constructs are important, because HCI research and psychology are often interested in the (causal) relationships between constructs (DeVellis, 2016). For instance, it may be helpful to study how motivation to engage with a product relates to experiences of frustration and usage frequency. In this situation, the user's specific pattern of motivational regulation – a construct – is of interest and not responses to single items. Furthermore, measures directly derived from a theory have numerous advantages, because theories offer definitions of constructs and formalized relationships and structures that are testable with empirical data. Thus, the items or indicators of theoretical constructs can be derived from these definitions. In turn, findings based on such a scale can inform theory and contribute to the aggregation of knowledge. However, in some situations a theoretical model is not readily available and more practical considerations are the focus. This is probably a common situation in HCI, and might be one of the leading causes for the widespread use of ad-hoc measures (Bargas-Avila & Hornbæk, 2011).

However, as suggested by manuscript 2, a lack of a theoretical model must not hinder the development of valid measures. The development of the LQS followed a bottom-up strategy in a situation without a theoretical framework. However, in contrast to ad-hoc scales, the

LQS has undergone validation. The items of the LQS in manuscript 2 can be understood as direct indicators of an atheoretical measure. Single items of the LQS may be examined and used as direct measures of quality for a particular aspect. Moreover, a participant's average agreement on the LQS can be understood as an indicator of perceived language quality. In the latter situation, researchers explicitly or implicitly assume an underlying model (meaning responses to the items are dependent on the construct "perceived language quality"). This shows that distinguishing between theoretical and atheoretical measures is sometimes difficult. However, if researchers are interested in the more general concept of perceived language quality, it is important to study the correlational structure of the scale and to conduct more advanced psychometric analyses. The exploratory factor analysis in manuscript 2 revealed that the items of the LQS do not measure one construct, but rather suggest that users actually judge text quality of an interface on two related dimensions, which we labeled linguistic correctness and readability. In the context of manuscript 2, this observation was not regarded as problematical, because user responses to the items of the LQS are meaningful to inform designers and product managers. However, from the perspective of theory building, multidimensionality of a construct is an essential observation.

For instance, the GEQ (a popular questionnaire measuring player experience), was developed following a bottom-up strategy (Poels et al., 2007). The authors of the GEQ assumed that it measures seven dimensions (or constructs) of the player experience. However, despite its widespread use, the authors have never formally evaluated this assumption. Our research indicated that the GEQ does not satisfy the criterion of structural validity; therefore, it should not be used in its current form (Brühlmann & Schmid, 2015; Law et al., 2018). Hence, it is vital to examine the scales' psychometric properties thoroughly when multiple items are thought to measure a hypothetical construct. Scales that reveal other dimensions than expected might challenge the items or the definition of the construct. Therefore, the relevance of linguistic correctness and readability (and their correlation) are subjects for further research, and could inform theory building around perceptions of product (language) quality.

Several factors differentiate the LQS from simple ad-hoc measures: its item properties have been examined in diverse situations (such as different languages and products), it has known values of reliability, and (most importantly) its validity has been studied. Besides the important contribution of theory- and model-based scales, validated scales developed with a bottom-up strategy also have various favorable properties. Measurement and quantification of UX are useful, because they allow comparisons of different products (Law et al., 2009), and facilitate the study of causes and effects of positive or negative experiences.

Taken together, in each specific area of study, manuscripts 1, 2, and 3 contribute to research by providing new or improved measurement tools. Manuscript 1 reports the de-

velopment and validation of a scale measuring user motivation as understood in SDT. Additionally, for the UMI, the rich body of evidence in SDT about how different motivational regulations relate to people's experiences and behavior allows HCI researchers to postulate hypotheses in the context of experiences with technology. These hypotheses can then be pre-registered (see Cockburn et al., 2018) and tested, which may advance knowledge in HCI and hopefully lead to less fragmented research (Liu et al., 2014). Manuscript 3 offers researchers and practitioners a short and versatile semantic differential scale (the TrustDiff), which is a viable alternative to existing scales for user trust. Thus, manuscripts 1 and 3 provide a basis for studying theoretical models and their limits.

In some areas of HCI research, such as language quality, theories might not be readily available. In manuscript 2, the development and validation of a scale without an overarching theory is reported, demonstrating it is still useful to develop measures following best practice to avoid the drawbacks of ad-hoc scales. Furthermore, thorough development and validation of in principle atheoretical measures may inform theory-building.

Finally, in manuscript 1 the development and validation of the UMI are reported in great detail, and all materials, data, and analysis scripts used for the validation are published online. Thus, this manuscript contributes to the call for increased transparency in research (Kay et al., 2016), and hopefully provides a template for future scale-development endeavors in HCI. This thesis addresses the issue of theory-driven research in HCI from a measurement perspective and demonstrates how profoundly intertwined theory and measurement are (Bringmann & Eronen, 2016).

Careless responding and online research

Even when the validity of a questionnaire has been established, other threats to validity may occur in HCI research. Construction of good surveys is both an art and a science (Fowler, 2013), and many pitfalls in design and sampling need to be avoided (Brühlmann & Mekler, 2018). For instance, Ludeke and Larsen (2017) found various data quality and corresponding problems in the Big Five scale from a recent World Values Survey. My analysis of the World Values Survey data revealed that some samples from specific countries have high rates of carelessness, as measured by the LongString analysis (Brühlmann, 2017). Studies conducted online are particularly susceptible to such problems with data quality. The conclusions drawn from studies with low data quality may be unreliable or even false; thus, of low validity. There are different forms of bad data, but the phenomenon of carelessness has received increased interest from the research community. Manuscript 4 demonstrated that carelessness, as measured by various indicators, is frequent in crowdsourced studies, providing researchers with a set of measures that should help to identify such data. It also showed that the rate of carelessness depends on the types of methods applied in a study,

and may even change throughout a survey. Initially, it may seem ironic that manuscript 4 describes how data quality on crowdsourced platforms is often of questionable quality, while the studies reported in manuscripts 1 and 3 used crowdsourced samples. However, both studies reported in manuscript 1 were conducted on MTurk, and manuscript 3 reports on two studies that were conducted on MTurk and a study conducted on FigureEight. Although not without criticism, data quality on MTurk is often better than on FigureEight (Peer et al., 2017) and all studies employed several data quality checks. Furthermore, the studies reported in manuscript 2 indicate that even self-selected samples are prone to data quality problems, and several steps are needed to ensure high-quality data.

Based on manuscript 4, several recommendations can be given to researchers. Data quality needs to be assessed in every online study (ideally with multiple measures), to avoid problems with scales or false conclusions. Further, a task-specific measure (such as the quality of a free-text answer) is essential for assessment. Ideally, this should be used in combination with attention checks and an investigation of answer patterns with the LongString index. The high proportion of carelessness detected in the study of manuscript 4 might persuade researchers to question online samples entirely. However, with the need for larger samples, coupled with the fact that a large portion of people's time is actually spent online, it makes sense from a perspective of efficiency, statistical power, and validity to conduct research online.

The studies reported in this thesis will help in conducting more rigorous and replicable research in HCI, by providing validated measures for three different domains and methods to ensure data quality. In combination with a solid methodology, valid measures and valid data are the basis of good scientific research. Together, these manuscripts contribute to the important issues of online survey research (Brühlmann & Mekler, 2018).

Limitations and future directions

Manuscript 1 reports the theory-grounded development of the UMI: a multidimensional measure of user motivation to use a specific technology. The reported development and validation studies entail initial and thorough validation. The UMI is rooted in SDT, but several other aspects of SDT need to be investigated in technology use, and how it may relate to existing models such as Hassenzahl et al. (2010) and the Technology Acceptance Model (Venkatesh, Morris, Davis, & Davis, 2009). These two models understand interaction with technology differently (Hornbæk & Hertzum, 2017) but both incorporate some aspects of SDT, such as need fulfillment or intrinsic and extrinsic motivation. Additionally, SDT has been successfully applied in the domain of work motivation (Gagné et al., 2015). Hence, the UMI provides a suitable framework for studying user motivation in this context. Accordingly, manuscript 1 provides a starting point for future research on user motivation

and its effects on positive or negative experiences with technology. From a scale development perspective, it is important to note that the two studies reported in manuscript 1 are not a finalized validation of the scale. It is the domain of future research to explore scale properties with different samples and specific contexts, and to study how the UMI relates to user behavior. The detailed and elaborate validation of the VisAWI could serve as a model (Moshagen & Thielsch, 2010). However, the psychometric properties of the UMI are promising and the UMI has been examined with two large samples and various products.

While manuscript 1 focuses more directly on theoretical implications, manuscripts 2 and 3 were also motivated by practical issues around measuring language quality (LQS) and trust (TrustDiff). Manuscript 2 describes in great detail the development of a tool to evaluate interface language quality, that may help to improve translation quality at a lower cost. While the LQS creates an idiosyncratic construct of language quality, special care was taken to ensure that the items of the LQS can be readily applied with different types of interfaces and in various contexts. However, the LQS has not been tested with a more diverse set of products and on mobile platforms. From a theoretical perspective, the two factors of linguistic correctness and readability warrant further research. While a more in-depth analysis of differences in specific languages was not in the scope of manuscript 2, the LQS might provide further insight into the particular challenges of translating measures. The psychometric properties of all translations were satisfactory, but there might be culture- and language-specific differences in the perception of items or rating scales that could distort results. For instance, in specific contexts, ratings may be systematically lower or higher because of cultural differences in acquiescence, or because of poor translation (Heine, Lehman, Peng, & Greenholtz, 2002). Such biases render cross-country comparison of means for individual items (or an overall score) difficult. Therefore, more research is needed to interpret LQS statistics with data from users of different cultural backgrounds correctly. However, the LQS provides a solid basis to study these phenomena and to develop more elaborate models of product quality perception.

The development of the TrustDiff in manuscript 3 followed the three-factor model of trust (e.g., McKnight et al., 2002b). Results of the confirmatory factor analysis in study 2 demonstrated that the data collected by the items of the TrustDiff follow this structure. In contrast to the UMI and the LQS, the TrustDiff adopts the form of a semantic differential scale. Semantic differential scales, such as the AttrakDiff (Hassenzahl et al., 2010), are influential in UX research and have several advantages over traditional agreement scales (Verhagen et al., 2015). Considering the interpretation of scores from different cultures, research hints that semantic differentials may be superior to Likert-type scales in cross-cultural settings (Maclay & Ware, 1961). As with the UMI and the LQS, validation of the TrustDiff scale is not complete. Future research should investigate whether the three-factor

model of trust is appropriate, and if user scores on these three dimensions are correlated with trust-related behaviors (such as number of purchases, or the frequency of website visits). However, the three studies reported in manuscript 3 with a total of over 1000 participants show that the TrustDiff has favorable psychometric properties and establish the scale as a valuable alternative to existing scales.

Regarding data quality and validity of data in online surveys, it is important to consider that manuscript 4 only examined FigureEight, one of several platforms where online studies can be conducted. Thus, the estimate that approximately half of the participants are careless may not readily transfer to other platforms or contexts. It remains challenging to select appropriate methods to detect carelessness and to define meaningful cut-offs for methods such as the LongString index, because they are study (or even scale) specific. All post-hoc detection methods in manuscript 4 were studied with the Big Five Inventory to ensure comparability with existing research (e.g., Meade & Craig, 2012). However, this constraint limits the analysis of carelessness in the study to this specific scale, and a particular place in the survey. A careless response is defined as being unrelated to the content of the item, but it is conceivable that the presentation of the questionnaire and the topic of the scale are a consideration. Motivation seems vital, and scales with odd wording or uninteresting topics might lead to inattention more quickly. Hence, future research should focus on one specific type of carelessness (such as patterned response), and study how survey design factors may reduce its occurrence. Another vital subject for future study is the unique requirements for surveys on mobile devices. Apart from new challenges concerning survey design and carelessness in this context, presentation and construction of scales might also be affected by smaller screens and new interaction modes (e.g., Couper, Antoun, & Mavletova, 2017).

The four manuscripts presented here make a substantial contribution to improving the problem of ad-hoc scales in HCI. They show how valid scales can be developed in three different domains with similar methods and different theoretical contexts. There is a call for more high-powered studies (e.g., Lakens & Evers, 2014), and questionnaire development is particularly dependent on large samples (DeVellis, 2016; Howard, 2016). Thus, better recruiting methods, data quality assurance, participant panels at universities, and more international collaboration are needed. There are already related new initiatives, such as the Many Labs project, which aims at replicating important findings in psychology and brings together researchers from around the world to collaborate in data collection (e.g., Buttrick et al., 2018). Additionally, the Psychological Science Accelerator, an international collaborative network of psychologists, collectively conduct studies in different countries (Moshontz et al., 2018). Similar efforts should also be undertaken by researchers in HCI to improve transparency and statistical power, and to promote replication.

Conclusion

The scale development studies in manuscripts 1, 2, and 3 report the development and validation of three thematically different measures with broad applicability regarding users and products following best practice. Even when these validated measures are employed, participant inattention can have problematic effects. In manuscript 4, new ways to study and ensure data quality are presented and evaluated in the context of crowdsourced online surveys. The manuscripts show that it is important to find an appropriate operationalization of constructs, but it is also vital not to take the measures as the only or the correct way to describe (or even define) concepts (Rosenberg, 2011). Therefore, it is beneficial to complement the self-reported data with behavioral measures, or follow a mixed-method approach in studying user experiences, which may reveal the limits of questionnaires (as in Petralito, Brühlmann, Iten, Mekler, & Opwis, 2017). However, these manuscripts provide a solid starting point and initial findings in their specific research topic areas that overcome the limits of ad-hoc, non-validated scales that are common in HCI (Bargas-Avila & Hornbæk, 2011). They offer a working definition (or a theoretically grounded operationalization of psychological constructs) that will enable researchers to investigate different phenomena in a way that may generalize over studies.

References

- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. doi: 10.3758/s13428-012-0265-2
- Bargas-Avila, J. A., & Brühlmann, F. (2016). Measuring user rated language quality: Development and validation of the user interface language quality survey (LQS). *International Journal of Human-Computer Studies*, *86*, 1–10. doi: 10.1016/j.ijhcs.2015.08.010
- Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11* (pp. 2689–2698). New York, NY, USA: ACM. doi: 10.1145/1978942.1979336
- Bart, Y., Shankar, V., Sultan, F., & Urban, G. L. (2005). Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study. *Journal of Marketing*, *69*(4), 133–152. doi: 10.1509/jmkg.2005.69.4.133
- Beck, A. T., Ward, C. H., Mendelson, M. M., Mock, J. J., & Erbaugh, J. J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561–571. doi: 10.1001/archpsyc.1961.01710120031004
- Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011, Mar 25). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. doi: 10.3758/s13428-011-0081-0

- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. doi: 10.1037/a0021524
- Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). Mmpi-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340–345. doi: 10.1037//1040-3590.4.3.340
- Bhattacharjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems*, *19*(1), 211–241. doi: 10.1080/07421222.2002.11045715
- Birk, M. V., Atkins, C., Bowey, J. T., & Mandryk, R. L. (2016). Fostering intrinsic motivation through avatar identification in digital games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2982–2995). New York, NY, USA: ACM. doi: 10.1145/2858036.2858062
- Blais, M. R., Sabourin, S., Boucher, C., & Vallerand, R. J. (1990). Toward a motivational model of couple happiness. *Journal of Personality and Social Psychology*, *59*(5), 1021–1031. doi: 10.1037/0022-3514.59.5.1021
- Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. Berlin: Springer Science & Business Media. doi: 10.1007/0-387-28981-X
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061–1071. doi: 10.1037/0033-295X.111.4.1061
- Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*, *26*(1), 27–43. doi: 10.1177/0959354315617253
- Brühlmann, F. (2017, March 23). *Can we trust big five data from the WVS?* [Blog]. <https://bruehlmann.io/blog/dataquality/2017/03/23/Can-we-trust-big-five-data/>.
- 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: Oxford University Press. doi: 10.1093/oso/9780198794844.003.0009
- Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2018). *Half of the participants in online surveys respond carelessly: An investigation of data quality in crowdsourced samples*. (Manuscript submitted for publication)
- Brühlmann, F., Petralito, S., Rieser, D. C., Aeschbach, L. F., & Opwis, K. (2018). *Trustdiff: Development and validation of a semantic differential for user trust on the web*. (Manuscript submitted for publication)
- Brühlmann, F., & Schmid, G.-M. (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* (pp. 1181–1186). New York, NY, USA: ACM. doi: 10.1145/2702613.2732831
- Brühlmann, F., Vollenwyder, B., Opwis, K., & Mekler, E. D. (2018). Measuring the "why" of interaction: Development and validation of the user motivation inventory (UMI). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 106:1–106:13).

- ACM. doi: 10.1145/3173574.3173680
- Buttrick, N., Aczel, B., Aeschbach, L. F., Bakos, B. E., Brühlmann, F., Claypool, H., . . . Wood, M. (2018). *Many Labs 5: registered replication report of Vohs and Schooler (2008), study 1*. (Manuscript in preparation)
- Calvo, R. A., & Peters, D. (2012). Positive computing: technology for a wiser world. *Interactions*, 19(4), 28–31. doi: 10.1145/2212877.2212886
- Calvo, R. A., Peters, D., Johnson, D., & Rogers, Y. (2014). Autonomy in technology design. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14* (pp. 37–40). doi: 10.1145/2559206.2560468
- Casaló, L. V., Flavián, C., & Guinalú, M. (2007). The role of security, privacy, usability and reputation in the development of online banking. *Online Information Review*, 31(5), 583–603. doi: 10.1108/14684520710832315
- Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2156–2160. doi: 10.1016/j.chb.2013.05.009
- Chen, S. C., & Dhillon, G. S. (2003). Interpreting dimensions of consumer trust in e-commerce. *Information Technology and Management*, 4(2), 303–318. doi: 10.1023/A:1022962631249
- Chin, W. W., Johnson, N., & Schwarz, A. (2008). A fast form approach to measuring technology acceptance and other constructs. *MIS Quarterly*, 687–703. doi: 10.2307/25148867
- Cho, J. (2006). The mechanism of trust and distrust formation and their relational outcomes. *Journal of Retailing*, 82(1), 25–35. doi: 10.1016/j.jretai.2005.11.002
- Chuang, L. L., & Pfeil, U. (2018). Transparency and openness promotion guidelines for hci. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. SIG04:1–SIG04:4). New York, NY, USA: ACM. doi: 10.1145/3170427.3185377
- 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. 141:1–141:12). doi: 10.1145/3173574.3173715
- Comley, P. (2015). Online market research. In M. v. Hamersveld & C. d. Bont (Eds.), *Market Research Handbook* (pp. 401–419). Chichester, England: John Wiley & Sons Ltd. doi: 10.1002/9781119208044.ch21
- Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: A study of consumer perceptions. *Electronic Commerce Research and Applications*, 2(3), 203–215. doi: 10.1016/s1567-4223(03)00024-3
- Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In P. Biemer et al. (Eds.), *Total survey error in practice* (pp. 133–154). New York, US: Wiley.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. doi: 10.1016/j.jesp.2015.07.006
- Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227–268. doi: 10.1207/S15327965PLI1104_01
- del Galdo, E. M., & Nielsen, J. (Eds.). (1996). *International users interface*. New York, NY, USA: John Wiley & Sons, Inc.

- Deterding, S. (2016). Contextual autonomy support in video game play : A grounded theory. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 3931–3943). New York, NY, USA: ACM. doi: 10.1145/2858036.2858395
- DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Thousand Oaks, CA: Sage publications.
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71–75. doi: 10.1207/s15327752jpa4901_13
- Dix, A. (2017). Human–computer interaction, foundations and new paradigms. *Journal of Visual Languages & Computing*, 42, 122–134. doi: 10.1016/j.jvlc.2016.04.001
- Dogan, V. (2018). A novel method for detecting careless respondents in survey data: Floodlight detection of careless respondents. *Journal of Marketing Analytics*, 6(3), 95–104. doi: 10.1057/s41270-018-0035-9
- Driscoll, J. W. (1978). Trust and participation in organizational decision making as predictors of satisfaction. *Academy of Management Journal*, 21(1), 44–56. doi: 10.2307/255661
- Echtler, F., & Häussler, M. (2018). Open source, open science, and the replication crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. alt02:1–alt02:8). New York, NY, USA: ACM. doi: 10.1145/3170427.3188395
- Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22(5), 323 - 327. doi: 10.1016/j.intcom.2010.04.004
- Flavián, C., Guinalú, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, 43(1), 1–14. doi: 10.1016/j.im.2005.01.002
- Fowler, F. J. (2013). *Survey research methods*. Thousand Oaks, CA: Sage publications.
- Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, 40(5), 873–884. doi: 10.1016/j.paid.2005.08.015
- Fry, E. (1968). A readability formula that saves time. *Journal of Reading*, 11(7), 513–578.
- Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1631–1640). New York, NY, USA: ACM. doi: 10.1145/2702123.2702443
- Gagné, M., Forest, J., Vansteenkiste, M., Crevier-Braud, L., van den Broeck, A., Aspel, A. K., ... Westbye, C. (2015). The multidimensional work motivation scale: Validation evidence in seven languages and nine countries. *European Journal of Work and Organizational Psychology*, 24(2), 178–196. doi: 10.1080/1359432X.2013.877892
- Gefen, D. (2002). Reflections on the dimensions of trust and trustworthiness among online consumers. *ACM Sigmis Database*, 33(3), 38–53. doi: 10.1145/569905.569910
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 27(1), 51–90. doi: 10.2307/30036519
- Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, 66(1), 877–902. doi: 10.1146/annurev-psych-010814-015321

- Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (SIMS). *Motivation and Emotion, 24*(3), 175–213. doi: 10.1023/A:1005614228250
- Hassenzahl, M., Diefenbach, S., & Göritz, A. (2010). Needs, affect, and interactive products – facets of user experience. *Interacting with Computers, 22*(5), 353–362. doi: 10.1016/j.intcom.2010.04.002
- Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology, 25*(2), 91–97. doi: 10.1080/01449290500330331
- Hassenzahl, M., & Ullrich, D. (2007). To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers, 19*(4), 429–437. doi: 10.1016/j.intcom.2007.05.001
- Hawkins, D. I., Albaum, G., & Best, R. (1974). Stapel scale or semantic differential in marketing research? *Journal of Marketing Research, 11*(3), 318–322. doi: 10.2307/3151152
- Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What’s wrong with cross-cultural comparisons of subjective likert scales? the reference-group effect. *Journal of Personality and Social Psychology, 82*(6), 903–918. doi: 10.1037/0022-3514.82.6.903
- Hesse, B. W. (2018). Can psychology walk the walk of open science? *American Psychologist, 73*(2), 126. doi: 10.1037/amp0000197
- Hong, I. B., & Cho, H. (2011). The impact of consumer trust on attitudinal loyalty and purchase intentions in b2c e-marketplaces: Intermediary trust vs. seller trust. *International Journal of Information Management, 31*(5), 469–479. doi: 10.1016/j.ijinfomgt.2011.02.001
- Hornbæk, K., & Hertzum, M. (2017). Technology acceptance and user experience: A review of the experiential component in HCI. *ACM Transactions on Computer-Human Interaction, 24*(5), 33:1–33:30. doi: 10.1145/3127358
- Hornbæk, K., & Oulasvirta, A. (2017). What is interaction? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 5040–5052). New York, NY, USA: ACM. doi: 10.1145/3025453.3025765
- Hornbæk, K., Sander, S. S., Bargas-Avila, J. A., & Simonsen, J. G. (2014). Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3523–3532). doi: 10.1145/2556288.2557004
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674. doi: 10.1198/106186006X133933
- 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. doi: 10.1080/10447318.2015.1087664
- Howard, M. C., & Melloy, R. C. (2016). Evaluating item-sort task methods: The presentation of a new statistical significance formula and methodological best practices. *Journal of Business and Psychology, 31*(1), 173–186. doi: 10.1007/s10869-015-9404-y
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*(3), 828–845. doi:

- 10.1037/a0038510
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), 696–701. doi: 10.1371/journal.pmed.0020124
- Jarvenpaa, S. L., Tractinsky, N., & Saarinen, L. (1999). Consumer trust in an internet store: A cross-cultural validation. *Journal of Computer-Mediated Communication*, *5*(2), 0–0. doi: 10.1111/j.1083-6101.1999.tb00337.x
- 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. doi: 10.21236/ada388787
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi: 10.1177/0956797611430953
- John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (Vol. 2, pp. 102–138). New York, NY, US: Guilford Press.
- Johnson, D., Gardner, M. J., & Perry, R. (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*, 38–46. doi: 10.1016/j.ijhcs.2018.05.003
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. doi: 10.1016/j.jrp.2004.09.009
- Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. doi: 10.1177/1094428115571894
- Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon’s Mechanical Turk workforce. *Computers in Human Behavior*, *83*, 243–253. doi: 10.1016/j.chb.2018.02.005
- Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1105–1114). doi: 10.1145/2207676.2208557
- Kay, M., Haroz, S., Guha, S., & Dragicevic, P. (2016). Special interest group on transparent statistics in hci. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1081–1084). doi: 10.1145/2851581.2886442
- Kay, M., Haroz, S., Guha, S., Dragicevic, P., & Wacharamanotham, C. (2017). Moving transparent statistics forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17* (pp. 534–541). New York, NY, USA: ACM. doi: 10.1145/3027063.3027084
- Kim, Y., & Peterson, R. A. (2017). A meta-analysis of online trust relationships in e-commerce. *Journal of Interactive Marketing*, *38*(Supplement C), 44–54. doi: 10.1016/j.intmar.2017.01.001
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new

- readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Navy Research Branch Report*. doi: 10.21236/ada006655
- Koufaris, M., & Hampton-Sosa, W. (2004). The development of initial trust in an online company by new customers. *Information & Management*, 41(3), 377–397. doi: 10.1016/j.im.2003.08.004
- Lafrenière, M.-A. K., Verner-Filion, J., & Vallerand, R. J. (2012). Development and validation of the Gaming Motivation Scale (GAMS). *Personality and Individual Differences*, 53(7), 827–831. doi: 10.1016/j.paid.2012.06.013
- Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9(3), 278–292. doi: 10.1177/1745691614528520
- Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, 8(2), 142–164. doi: 10.1017/iop.2015.13
- 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. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM. doi: 10.31234/osf.io/u94qt
- Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 719–728). New York, NY, USA: ACM. doi: 10.1145/1518701.1518813
- Leiva, L. A., & Alabau, V. (2014). The impact of visual contextualization on UI localization. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3739–3742). New York, NY, USA: ACM. doi: 10.1145/2556288.2556982
- Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3553–3562). New York, NY, USA: ACM. doi: 10.1145/2556288.2556969
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584–585. doi: 10.1126/science.aal3618
- Lonsdale, C., Hodge, K., & Rose, E. A. (2008). The behavioral regulation in sport questionnaire (BRSQ): Instrument development and initial validity evidence. *Journal of Sport & Exercise Psychology*, 30(3), 323–355. doi: 10.1123/jsep.30.3.323
- Lu, J., Wang, L., & Hayes, L. A. (2012). How do technology readiness, platform functionality and trust influence C2C user satisfaction? *Journal of Electronic Commerce Research*, 13(1), 50–69.
- Ludeke, S. G., & Larsen, E. G. (2017). Problems with the big five assessment in the world values survey. *Personality and Individual Differences*, 112, 103–105. doi: 10.1016/j.paid.2017.02.042
- Maclay, H., & Ware, E. E. (1961). Cross-cultural use of the semantic differential. *Behavioral Science*, 6(3), 185–190. doi: 10.1002/bs.3830060303
- Mallett, C., Kawabata, M., Newcombe, P., Otero-Forero, A., & Jackson, S. (2007). Sport motivation scale-6 (SMS-6): A revised six-factor sport motivation scale. *Psychology of Sport and Exercise*,

- 8(5), 600–614. doi: 10.1016/j.psychsport.2006.12.005
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. doi: 10.1016/j.jrp.2013.09.008
- Markland, D., & Tobin, V. (2004). A modification to the behavioural regulation in exercise questionnaire to include an assessment of amotivation. *Journal of Sport and Exercise Psychology, 26*(2), 191–196. doi: 10.1123/jsep.26.2.191
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. doi: 10.2307/258792
- McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communications Monographs, 66*(1), 90–103. doi: 10.1080/03637759909376464
- McKnight, D. H., Choudhury, V., & Kacmar, C. (2002a). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334–359. doi: 10.1287/isre.13.3.334.81
- McKnight, D. H., Choudhury, V., & Kacmar, C. (2002b). The impact of initial consumer trust on intentions to transact with a web site: A trust building model. *The Journal of Strategic Information Systems, 11*(3), 297–323. doi: 10.1016/s0963-8687(02)00020-3
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. doi: 10.1037/a0028085
- Mekler, E. D., & Hornbæk, K. (2016). Momentary pleasure or lasting meaning? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (pp. 4509–4520). New York, NY, USA: ACM. doi: 10.1145/2858036.2858225
- Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market research relationships. *Journal of Marketing, 81*–101. doi: 10.2307/1252059
- Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und fragebogenkonstruktion*. Berlin: Springer. doi: 10.1007/978-3-642-20072-4
- Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies, 68*(10), 689–709. doi: 10.1016/j.ijhcs.2010.05.006
- Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., . . . Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1*–15. doi: 10.1177/2515245918797607
- Mullan, E., Markland, D., & Ingledew, D. K. (1997). A graded conceptualisation of self-determination in the regulation of exercise behaviour: development of a measure using confirmatory factor analytic procedures. *Personality and Individual Differences, 23*(5), 745–752. doi: 10.1016/S0191-8869(97)00107-4
- Muntés Mulero, V., Paladini Adell, P., España Bonet, C., & Màrquez Villodre, L. (2012). Context-aware machine translation for software localization. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation: EAMT 2012: Trento, Italy, May 28th-30th 2012* (pp. 77–80).
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . others (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. doi: 10.1126/

- science.aab2374
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi: 10.1126/science.aac4716
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. doi: 10.1016/j.jesp.2009.03.009
- Oulasvirta, A., & Hornbæk, K. (2016). Hci research as problem-solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4956–4967). New York, NY, USA: ACM. doi: 10.1145/2858036.2858283
- Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi: 10.1177/0963721414531598
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi: 10.1177/1745691612465253
- Pavlou, P. A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research*, *15*(1), 37–59. doi: 10.1287/isre.1040.0015
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. doi: 10.1016/j.jesp.2017.01.006
- Pelletier, L. G., Rocchi, M. A., Vallerand, R. J., Deci, E. L., & Ryan, R. M. (2013). Validation of the revised sport motivation scale (SMS-II). *Psychology of Sport and Exercise*, *14*(3), 329–341. doi: 10.1016/j.psychsport.2012.12.002
- Pelletier, L. G., Tuson, K. M., Green-Demers, I., Noels, K., & Beaton, A. M. (1998). Why are you doing things for the environment? The motivation toward the environment scale (MTES). *Journal of Applied Social Psychology*, *28*(5), 437–468. doi: 10.1111/j.1559-1816.1998.tb01714.x
- Pelletier, L. G., Tuson, K. M., & Haddad, N. K. (1997). Client motivation for therapy scale: A measure of intrinsic motivation, extrinsic motivation, and amotivation for therapy. *Journal of Personality Assessment*, *68*(2), 414–35. doi: 10.1207/s15327752jpa6802_11
- Petralito, S., Brühlmann, F., Iten, G., Mekler, E. D., & Opwis, K. (2017). A good reason to die: How avatar death and high challenges enable positive experiences. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 5087–5097). New York, NY, USA: ACM. doi: 10.1145/3025453.3026047
- Poels, K., de Kort, Y., & Ijsselstein, W. (2007). "It is always a lot of fun!": Exploring dimensions of digital game experience using focus group methodology. In *Proceedings of the 2007 Conference on Future Play* (pp. 83–89). New York, NY, USA: ACM. doi: 10.1145/1328202.1328218
- Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), 46–55.
- Rieser, D. C., & Bernhard, O. (2016). Measuring trust: The simpler the better? In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2940–2946). New York, NY, USA: ACM. doi: 10.1145/2851581.2892468

- Rosenberg, A. (2011). *Philosophy of science: A contemporary introduction*. Routledge. doi: 10.4324/9780203807514
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35(4), 651–665. doi: 10.1111/j.1467-6494.1967.tb01454.x
- Rozendaal, M. M. C., Keyson, D. V., & de Ridder, H. (2007). Product behavior and appearance effects on experienced engagement during experiential and goal-directed tasks. In *Proceedings of the 2007 conference on designing pleasurable products and interfaces* (pp. 181–193). doi: 10.1145/1314161.1314178
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57(5), 749–761. doi: 10.1037/0022-3514.57.5.749
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1), 54–67. doi: 10.1006/ceps.1999.1020
- Ryan, R. M., & Frederick, C. (1997). On energy, personality, and health: Subjective vitality as a dynamic reflection of well-being. *Journal of Personality*, 65(3), 529–565. doi: 10.1111/j.1467-6494.1997.tb00326.x
- Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, 30(4), 344–360. doi: 10.1007/s11031-006-9051-8
- Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Cambridge, MA: Morgan Kaufmann.
- Schlosser, A. E., White, T. B., & Lloyd, S. M. (2006). Converting web site visitors into buyers: How web site investment increases consumer trusting beliefs and online purchase intentions. *Journal of Marketing*, 70(2), 133–148. doi: 10.1509/jmkg.70.2.133
- Schriver, K. A. (1989, Dec). Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Transactions on Professional Communication*, 32(4), 238–255. doi: 10.1109/47.44536
- 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. doi: 10.1016/j.chb.2014.11.064
- Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, 80(2), 325–339. doi: 10.1037/0022-3514.80.2.325
- Sheldon, K. M., Ryan, R. M., Deci, E. L., & Kasser, T. (2004). The independent effects of goal contents and motives on well-being: It's both what you pursue and why you pursue it. *Personality and Social Psychology Bulletin*, 30(4), 475–486. doi: 10.1177/0146167203261883
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 0956797611417632. doi: 10.1037/e519702015-014
- Sun, H. (2001). Building a culturally-competent corporate web site: An exploratory study of cultural markers in multilingual web design. In *Proceedings of the 19th Annual International Conference on Computer Documentation* (pp. 95–102). New York, NY, USA: ACM. doi:

- 10.1145/501516.501536
- Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 271–360). San Diego, CA: Academic Press. doi: 10.1016/S0065-2601(08)60019-2
- Van Auken, S., & Barry, T. E. (1995). An assessment of the trait validity of cognitive age measures. *Journal of Consumer Psychology, 4*(2), 107–132. doi: 10.1207/s15327663jcp0402_02
- van Schaik, P., & Ling, J. (2009). The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies, 67*(1), 79 – 89. doi: 10.1016/j.ijhcs.2008.09.012
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2009). User acceptance of information technology: Toward a unified view. *Management Information Systems, 27*(3), 425–478. doi: 10.2307/30036540
- Verhagen, T., van Den Hooff, B., & Meents, S. (2015). Toward a better use of the semantic differential in is research: An integrative framework of suggested action. *Journal of the Association for Information Systems, 16*(2), 108–143. doi: 10.17705/1jais.00388
- Vlachopoulos, S. P., Katartzi, E. S., Kontou, M. G., Moustaka, F. C., & Goudas, M. (2011). The revised perceived locus of causality in physical education scale: Psychometric evaluation among youth. *Psychology of Sport and Exercise, 12*(6), 583–592. doi: 10.1016/j.psychsport.2011.07.003
- Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426–432. doi: 10.1037/a0022790
- Wilson, M. L., Mackay, W., Chi, E., Bernstein, M., Russell, D., & Thimbleby, H. (2011). RepliCHI – CHI should be replicating and validating results more: Discuss. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (pp. 463–466). doi: 10.1145/1979742.1979491
- Wirtz, J., & Lee, M. C. (2003). An examination of the quality and context-specific applicability of commonly used customer satisfaction measures. *Journal of Service Research, 5*(4), 345–355. doi: 10.1177/1094670503005004006

Acknowledgements

I am sincerely grateful to many people who supported me on my way to the completion of this thesis. Without their help this work would not have been possible:

- Klaus Opwis, my thesis supervisor, for offering me the opportunity to write this thesis and his encouragement, trust, and support over the last years.
- Elisa Mekler, head of the HCI research group, for teaching me the 'tools of the trade' of good research, for endless hours of engaged discussions, inspiration, and encouragement. Thank you so much!
- Javier Bargas-Avila, for volunteering to be Second Reviewer and for sparking my interest in HCI.
- My co-authors Klaus Opwis, Javier Bargas-Avila, Elisa Mekler, Serge Petralito, Beat Vollenwyder, Christoph Pimmer, Glena Iten, Lena Aeschbach, Denise Rieser, Alexandre Tuch, Gian-Marco Schmid.
- My former and current colleagues Markus Stöcklin, Alexandre Tuch, Silvia Heinz, Elisa Mekler, Livia Müller, Julia Bopp, Glena Iten, Sharon Steinemann, Serge Petralito, Beat Vollenwyder, Lars Frasseck.
- My assistants and students Lena Aeschbach, Denise Rieser, Yanira Gonzalez, Laura Quintana, Philipp Baumgartner.
- The doctoral committee for evaluating this work: Jana Nikitin (Chair), Klaus Opwis (First Reviewer) and Javier Bargas-Avila (Second Reviewer).
- Last but not least I want to thank my family who supported me even when times were not easy: Renate, Jürg, Viola, Warin, and most of all Mirjam. I love you.

curriculum vitae

Florian Brühlmann

Aus Datenschutzgründen entfernt

Appendix

1. **Brühlmann, F.**, Vollenwyder, B., Opwis, K., & Mekler, E. D. (2018). Measuring the “why” of interaction: Development and validation of the user motivation inventory (UMI). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. (pp. 106:1–106:13). New York, NY, USA: ACM.
doi: 10.1145/3173574.3173680
2. Bargas-Avila, J.A. & **Brühlmann, F.** (2016). Measuring user rated language quality: Development and validation of the user interface language quality survey (LQS). *International Journal of Human-Computer Studies*, 86, 1-10.
doi: 10.1016/j.ijhcs.2015.08.010
3. **Brühlmann, F.**, Petralito, S., Rieser, D. C., Aeschbach, L. F., & Opwis, K. (2018). *TrustDiff: Development and validation of a semantic differential for user trust on the web*. Manuscript submitted for publication.
4. **Brühlmann, F.**, Petralito, S., Aeschbach, L. F., & Opwis, K. (2018). *Half of the participants in online surveys respond carelessly: An investigation of data quality in crowdsourced samples*. Manuscript submitted for publication.

Measuring the “Why” of Interaction: Development and Validation of the User Motivation Inventory (UMI)

Florian Brühlmann, Beat Vollenwyder, Klaus Opwis, Elisa D. Mekler

Center for Cognitive Psychology and Methodology, Department of Psychology, University of Basel
{florian.bruehlmann, beat.vollenwyder, klaus.opwis, elisa.mekler}@unibas.ch

ABSTRACT

Motivation is a fundamental concept in understanding people’s experiences and behavior. Yet, motivation to engage with an interactive system has received only limited attention in HCI. We report the development and validation of the User Motivation Inventory (UMI). The UMI is an 18-item multidimensional measure of motivation, rooted in self-determination theory (SDT). It is designed to measure intrinsic motivation, integrated, identified, introjected, and external regulation, as well as amotivation. Results of two studies (total $N = 941$) confirm the six-factor structure of the UMI with high reliability, as well as convergent and discriminant validity of each subscale. Relationships with core concepts such as need satisfaction, vitality, and usability were studied. Additionally, the UMI was found to detect differences in motivation for people who consider abandoning a technology compared to those who do not question their use. The central role of motivation in users’ behavior and experience is discussed.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI); Miscellaneous; J.4. Social and Behavioral Sciences: Psychology

Author Keywords

Motivation; user experience; technology use; self-determination theory; scale development; usability

Open data policy

All data, study materials, and main analysis code used for the development and validation of the UMI are available at <https://osf.io/m3fbk/>

INTRODUCTION

Motivation is a fundamental concept in our lives, as it drives all intentional behavior. This also holds true for technology use since the motivation to engage with a given interactive system is at the core of the formation of user experience [22]. The reasons why people engage with a technology affect how users perceive product qualities, what qualities are important, and how they affect the users’ experience. For instance, the

pursuit of instrumental, task-directed goals renders usability problems more salient, which may in turn negatively influence users’ experience and retrospective product evaluation [20]. In contrast, non-utilitarian qualities are preferred when pursuing more exploratory [56] or experientially motivated behaviors involving technology use (e.g., to have fun, [20, 41]). Product beauty was found to be more important for leisurely rather than work-related technology use [52]. Similarly, users’ experiences with technology varied when pursued for eudaimonic (e.g., developing one’s personal potential) or for hedonic (e.g., pleasure) reasons [31]. In fact, depending on users’ motivation, the same technology-supported activities might be experienced very differently, say, when playing digital games for leisurely or for professional purposes [11].

Self-determination theory (SDT), an influential theory of human motivation [10, 43], differentiates the *what* (i.e., goal content) and the *why*, that is, the regulatory processes underlying goal pursuit [8]. According to SDT, people can satisfy the innate psychological needs for autonomy, competence, and relatedness [8] through a variety of behaviors. However, the quality of people’s behavior, the extent of need satisfaction, as well as the consequences on well-being depend on the motivational regulations underlying these behaviors [8]. Surprisingly, while need satisfaction – itself a key concept of SDT – was repeatedly found to be a defining characteristic of positive user experience (e.g., [19, 31]) and considered core to the understanding of what makes interaction good [22], the regulatory processes posited by organismic integration theory (OIT), a sub-theory of SDT, have so far received scant attention within HCI research – which may in part be due to the lack of a suitable measuring instrument. Distinguishing between different regulations might provide a more nuanced understanding of positive (and negative) experiences and their effects on need satisfaction (e.g., [58]). Additionally, motivation is a fundamental element to consider in studies concerning the effects of technology on well-being [31]. Hence, a multidimensional scale of motivation could extend existing models of user experience [27] and complement qualitative approaches to the “why” of interaction by providing a reliable tool that can be used to find generalizable and replicable results. A questionnaire for motivation can be applied to test theories and hypotheses and establish causal relationships in randomized controlled experiments.

In the present work, we describe the development and validation of the User Motivation Inventory (UMI). Our contribution is three-fold: First, the results of two validation studies (total sample size $N = 941$) indicate that the UMI has excellent psychometric properties, measuring six different types of mo-



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI 2018 April 21–26, 2018, Montreal, QC, Canada

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5620-6/18/04.

DOI: <https://doi.org/10.1145/3173574.3173680>

Differing consequences of motivational regulations

The different types of motivations on the self-determination continuum are associated with different behavioral, cognitive and emotional consequences. SDT postulates that the consequences are decreasingly positive the less self-determined the quality of a behavior is [10]. Specifically, intrinsic motivation is expected to lead to the most positive consequences, followed by integrated and identified regulation, which are forms of extrinsic motivation. Introjected and external regulation are presumed to lead to negative consequences, and amotivation to the most negative consequences [10]. Several studies have provided evidence that more self-determined types of motivation (intrinsic motivation, integrated regulation, and identified regulation) lead to the most positive behavioral (e.g. greater persistence), cognitive (e.g., enhanced concentration), and emotional (e.g., more positive emotions, greater well-being) outcomes when compared to nonself-determined types of motivation (introjected regulation, external regulation, and amotivation) [8, 10]. Support for this notion has been found for a wide variety of life domains such as academic achievement [17], sports [29, 33, 37], romantic relationships [3, 16], environmental protection [38], therapy motivation [39], and consumer behavior [58]. Zhang et al. [58], for instance, found that while experiential purchases (e.g., holidays) are typically regarded as more positively related to well-being than material purchases, this effect largely depended on people's motivational regulation. People who spent money on experiential purchases for autonomous reasons, meaning that they regarded them as an important part of their life, reported more need satisfaction, more flourishing, and vitality than people who spent money on these experiences for controlled reasons, such as for the recognition they got from others.

As with other life domains, technology use is likely motivated by different regulations. For instance, the notions of hedonic and eudaimonic motivation employed in the study of Mekler and Hornbæk [31] bear much semblance to intrinsic motivation and integrated regulation respectively, but do not account for less autonomous regulations. In another example, LaFrenière et al. [26] followed OIT when developing a scale for assessing gaming motivation and also found that more autonomous regulations (i.e., intrinsic motivation, integrated and identified regulation) were associated with need satisfaction, while the less self-determined regulations (i.e., introjected and external regulation, amotivation) were not. However, due to their instrument being specific to (arguably certain) games only (items include e.g., "I play video games to acquire powerful rare items"), it is not readily applicable to assessing people's motivation for using other interactive systems.

Yet given the great influence these different regulations might have on people's experience and use of interactive technology, a better understanding of users' motivation – the *why* of interaction – is imperative. To this end, we aim to measure different types of motivation for technology use. In the first study, a new measure of user motivation is developed. In the second study the underlying theoretical structure is verified and the impact of different types of motivation on usability, well-being and likelihood to recommend is investigated.

Development and validation strategy

The development and validation of the UMI followed best practices [12, 32]. In the first phase, we reviewed existing scales and adapted items to reflect the theoretical dimensions of motivational regulation in the context of technology use. This large item pool was subject to an item sort task and further refinement by the authors. This phase also included an independent expert review of content validity. In the second phase, the item pool was administered to a development sample in Study 1 to optimize scale length and identify the best items reflecting each of the six motivational regulations. In the third phase, we explored the dimensionality, reliability, convergent and discriminant validity with an independent validation sample in Study 2. To ensure construct validity, we also studied relations of these six motivational regulations to conceptually relevant measures from SDT and UX research. In the fourth phase, we investigated how motivational regulation differs in people who had thought about abandoning a technology compared to those who never thought about quitting.

ITEM POOL DEVELOPMENT AND REVIEW

Existing scales

In line with previous SDT research, the UMI was designed to measure the general motivation to engage with a specific technology. Existing scales on motivation were the basis of item development along with the definition of the different types of regulation described in the Handbook for Self-Determination Research [9]. The existing scales we used as item sources were developed for the areas of academic achievement (SIMS [17]), video games (GAMS, [26]), sports (BRSQ [28]; BREQ [34] and BREQ-2 [30]; SMS-6, [29]; SMS-II, [37]; PLOC-R, [57]), environmental protection (METS, [38]), romantic relationships (CMQ, [3]), therapy motivation (CMOTS, [39]), school (PLOC, [42]), and well-being ([50]). While some motivational regulation scales do not include all dimensions posited by organismic integration theory (e.g., SIMS [17] or BREQ-2 [30]), we opted to cover all six dimensions to adequately represent the theoretical foundation and granularly differentiate between all regulations. Particular care was therefore taken to ensure that the UMI items have as little overlap as possible. Still, we expected that the items for these facets will correlate more strongly the closer to one another they are on the spectrum of motivation [42]. Based on these scales, an initial item pool of 249 items was created by the first author. Particular care was taken to adapt the wording of the items to reflect technology use. In a next step, all authors reviewed the item pool and removed or rephrased duplicates, near-duplicates, as well as items that were too specific. A pool of 150 items remained, which were, similar to the User Burden Scale [51], rephrased to include a placeholder for the technology in question (e.g., "I enjoy using [x]").

Item Sort Task

The first, second and last author independently conducted an item sort task [24] with all 150 items. Any items that did not receive a 100% agreement on the intended construct were removed, unless one of the authors involved at this stage vetoed the removal of an item. A total of 102 items remained.

Expert review

Two psychologists with expertise in self-determination theory, but who were not themselves involved in this research project, reviewed the pool in a 2-hour workshop. The goal was to review content validity, that is, ensure that all relevant aspects of motivation for technology use were covered. The experts rated each item on relevance, clarity and checked whether any aspects were missed. This review led to the removal of fifteen items and rewording of four items. Additionally, six items capturing integrated regulation were created. At this stage, the questionnaire consisted of 93 items. Amotivation was measured with 16 items, external regulation with 12, introjected regulation with 18, identified with 12, integrated regulation with 20, and intrinsic motivation with 15 items.

STUDY 1

The purpose of Study 1 was to test and reduce the UMI item pool to identify the best items measuring the proposed six regulations. To this end, we deployed a survey on Amazon Mechanical Turk and conducted an analysis with four steps: 1) psychometric analysis of all 93 items; 2) factor analysis for a subset of items; 3) selection of the best items; 4) factor analysis with these items for a preliminary structural validation. The number of items and the expected communalities determine the sample size required for factor analysis [12]. In general, a sample size of at least 200 participants is recommended [23]. We expected high communalities and good performance of the items as they were based on existing scales. Nevertheless, since the number of items under examination was large, we aimed for a sample size of over 450.

Procedure

After providing consent, participants were asked to fill in basic demographic information (gender, age, and experience with games). Next, they named an interactive technology that they used frequently. The focus was set on a frequently used technology to make sure that the UMI is applicable to widely used technologies. Participants were then asked to describe the technology and explain how they use it. The rationale behind this question was that if an uncommon website or technology was listed, we would have some information about what it is and how it can be used. We then also asked participants to report how frequently the technology was used and asked them to answer several scales in relation to this specific technology that will be discussed in the next paragraph. On the last page, participants were asked to indicate whether they answered the questions seriously (this served as a self-reported measure of data quality), they also had possibility to comment on the study and were given a completion code for Mechanical Turk. An instructed response item was included in the UMI items to filter out careless participants.

Participants

A total of 507 participants from the US completed the full questionnaire on Amazon Mechanical Turk. The survey took 13 minutes ($SD = 5.8$ minutes) to complete on average.

Data cleaning

Based on the recommendations by [7], 17 participants were excluded because they completed the survey in less than 4

minutes or not in one session. Two additional participants were excluded because they selected the same answer for more than 83.3% of the UMI items (5 out of 6). Seven participants were excluded because of a negative person-total correlation, which is an indicator for very unusual answering patterns [7].

Sample description

A total of 481 participants ($M_{age} = 38.31, SD = 12.61, range = 19 - 75$; 39.1% male, 1.5% non-binary or not specified) were included in the analysis. Participants could freely report on any interactive technology they used frequently. A majority of 33% chose Facebook, 11% a not further specified Smartphone, 10% iPhone, and 46% various other technologies, such as the Fitbit, other handheld devices such as Android OS, iPad, video game consoles such as the Playstation 4, as well as other social networking services such as Reddit and Twitter. With regards to the last 14 days, 45.9% of participants indicated that they used the interactive technology on average six times a day or more, 17.5% four to five times per day, 21.2% two to three times per day and 15.4% once a day or less.

Measures

In addition to the UMI items and demographic variables, five scales were included in Study 1. However, due to space concerns, only the measures relevant for the development of the UMI are reported here. The other scales were also included in Study 2 and are discussed in more detail in the Measures section of Study 2. Please note that all measures and data from *both* studies are available on <https://osf.io/m3fbk/>.

Type of technology

Participants could name any single technology. The statistical software R was used to semi-manually clean this data to ensure that typos and different spellings were associated with the correct technology name.

Frequency of use

A single-item measure captured frequency of use: *How frequently did you use this [referring to the technology they named above] interactive technology in the last 14 days?*

UMI

The 93 items of the initial UMI item pool were distributed over four pages and displayed in random order. A 7-point Likert-type scale ranging from 1 (strongly disagree) to 7 (strongly agree) was used. Two attention check items were implemented to flag and exclude participants that answered carelessly: *This is a verification item, please select strongly disagree* and *I read instructions carefully. To show that you are reading these instructions, please leave this question blank*.

Results

The analysis was twofold: A psychometric item analysis was performed to remove problematic items, followed by an exploratory factor analysis to examine the structure and reduce the number of indicators to a scale with 3 items per construct.

Item analysis

Since the UMI was intended to measure six distinct (but possibly correlated) constructs based on SDT, it was decided to

investigate descriptive statistics, difficulty indices, item variance, and discriminatory power for each construct separately. Please refer to the online materials for statistics on all 93 items. The item analysis followed recommendations by Moosbrugger and Kelava [32]. We removed one item with a variance of less than 1, because items with low variance are not suitable for differentiating between participants [32]. Discriminatory power describes how a single item's ability to differentiate between participants relates to the ability of the entire scale to differentiate between participants. Two items were removed because their discriminatory power was below the recommended value $< .30$ [4]. For each construct, inter-item correlations and average inter-item correlation (homogeneity) were investigated. Six items with an average inter-item correlation of less than $.4$ were removed, resulting in 84 items remaining for factor analysis.

Exploratory Factor Analysis (EFA)

Bartlett's test indicated factorability ($\chi^2_{df=3486} = 34439.59, p < .001$) as did the average Kaiser-Meyer-Olkin factor adequacy measure (Overall MSA = 0.96, none below .8). Parallel analysis suggested seven factors but the visual inspection of the Scree plot and our theoretical assumption suggested six factors. Since the goal was to reduce the number of items of the UMI and construct a scale that is consistent with the theoretical model, we conducted the EFA with six factors.

Data were tested for multivariate normality with Mardia Tests ($\chi^2_s = 181720.4, p < .001$; $Z_k = 127.4, p < .001$). Both tests indicated highly non-normal data. Hence, we chose to use principal axis factoring. The six factors had eigenvalues greater than 1.99 and explained 59% of the total variance. The factors were rotated using the Oblimin method, because following theory some of the factors were expected to be correlated. Results revealed factor correlations ranging from $-.31$ to $.56$. The factor loadings and communalities of all 84 items are reported in the online materials. The communality of one item was below the recommended threshold of $.30$ and subsequently removed. To further reduce the item pool and improve the measurement model, items were removed based on two criteria. First, items that showed substantial ($>.30$) cross-loadings were removed. Second, items with the highest loadings on the primary factor were retained unless there was already an item selected with very similar wording. The goal was to balance psychometric and theoretical grounds for item selection and optimize scale length [12]. Based on these criteria, the best eighteen items with three items per factor were identified (item wording depicted in Table 1).

These eighteen items were then subject to a second analysis. Bartlett's test was significant ($\chi^2_{df=153} = 5238.14, p < .001$) and the average Kaiser-Meyer-Olkin measure was $.83$ with none below $.75$. Parallel analysis and inspection of the Scree plot suggested six factors and data were non-normal (Mardia tests: $\chi^2_s = 6270.91, p < .001$; $Z_k = 59.75, p < .001$).

Principal axis factoring (Oblimin rotation) extracted six factors with eigenvalues greater than 0.89 explaining 71% of the total variance. Correlations of the six factors and internal consistencies are shown in Table 2. Factor loadings and com-

munalities of the final 18 items are depicted in Table 1. All eighteen items loaded substantially on their designated factor without any notable cross-loadings (none above $.3$), reflecting the theoretically assumed six-factor structure.

Discussion

In Study 1, we identified eighteen items measuring six related constructs. In line with OIT, the correlations between the factors show that conceptually close regulations such as intrinsic motivation and integrated regulation correlate more strongly than intrinsic motivation and introjected regulation. The results of the second factor analysis support a six dimensional measure with high reliability and good psychometric properties. This structure was put to test and investigated in relation to other measures in Study 2. As with Study 1, we aimed at a sample size of over 450 participants.

STUDY 2

The goal of Study 2 was to validate the measurement model with a different sample applying confirmatory factor analysis (CFA). To test construct validity, we first investigated fit measures of the proposed model and compared them to alternative models. Second, to ensure convergent and divergent validity, we studied how the different UMI dimensions relate to other relevant constructs from SDT and user experience research. Third, motivational patterns between participants who questioned their use and those who never did so were investigated to gain a deeper understanding of the interplay between users' intentions, behavior and motivation.

Procedure

Study 2 was largely patterned after Study 1. After providing consent and basic demographic information, participants were again asked to name an interactive technology that they used frequently, followed by the eighteen UMI items, four additional open questions and the remaining measures.

Participants

A total of 498 participants from the United States completed the full survey on Amazon Mechanical Turk. The survey took 8 minutes ($SD = 4.4$ minutes) to complete on average. None had previously partaken in Study 1.

Data cleaning

Data were cleaned following the same procedure from Study 1. Five participants were excluded because they listed more than one technology or did not comply with the task. One participant indicated that we should not use their data. Four participants completed the survey in less than 3 minutes or not in one single session. One participant selected the same answer for all 18 items of the UMI and 27 participants had a negative person-total correlation.

Sample description

After data cleaning, a total of 460 participants ($M_{age} = 37.38, SD = 11.74, range = 18 - 76$; 40.4% male, less than 1% non-binary or not specified) were included in the analysis. Again, participants could freely choose any interactive technology they used frequently. A majority of 42% chose Facebook, 5% Instagram, 5% Twitter, 4% Fitbit, and 44% various other

Subscale	Item	Factor						h ²
		AMO	EXT	INJ	IDE	INT	IMO	
Amotivation	1. I use [X], but I question why I continue to use it	.864						.739
	2. I use [X], but I wonder what is the point in using it	.854						.739
	3. I use [X], but I don't see why I should keep on bothering with it	.830						.711
External regulation	1. Other people will be upset if I don't use [X]		.799					.571
	2. I use [X] because others will not be pleased with me if I don't		.836					.727
	3. I feel under pressure from others to use [X]		.736					.681
Introjected regulation	1. I would feel bad about myself if I quit [X]			.943				.855
	2. I would feel guilty if I quit using [X]			.840				.740
	3. I would feel like a failure if I quit using [X]			.827				.728
Identified regulation	1. Using [X] is a sensible thing to do				.546			.458
	2. The benefits of using [X] are important to me				.742			.601
	3. Using [X] is a good way to achieve what I need right now				.872			.744
Integrated regulation	1. I use [X] because it reflects the essence of who I am					.795		.714
	2. Using [X] is consistent with my deepest principles					.773		.695
	3. I use [X] because it expresses my values					.932		.818
Intrinsic motivation	1. I use [X] because it is enjoyable						.849	.734
	2. I think using [X] is an interesting activity						.760	.637
	3. Using [X] is fun						.929	.843

Table 1. Pattern matrix from the EFA in Study 1 ($N = 460$) with the final version of the UMI. Loadings of all 18 items on the six factors are depicted, loadings below .3 are not shown. h^2 = Communalities. [X] is a placeholder for the technology chosen by the participants.

	1.	2.	3.	4.	5.	α
1. AMO	-					.89
2. EXT	.30	-				.84
3. INJ	.14	.44	-			.91
4. IDE	-.21	.11	.33	-		.80
5. INT	.02	.19	.50	.47	-	.89
6. IMO	-.33	-.20	.06	.15	.40	.89

Table 2. Factor correlations and internal consistency (Cronbach's α) for Study 1 with 18 items.

technologies. Among them were productivity software such as MS Word or Excel, other social media networks such as YouTube and Reddit, as well technologies such as Amazon Echo or Android OS. Over the last 14 days, 38.3% of the participants indicated that they used the interactive technology on average six times a day or more, 19.8% four to five times per day, 22.2% two to three times per day and 19.8% once a day or less.

Open questions about technology use

After answering the UMI, participants were asked to describe in their own words why they use the technology in question (1). For illustration purposes, two contrasting answers from two different participants about the same technology are presented:

I use Facebook because it is a way to connect to people [...] using Facebook allows me to see family photos, to reach out to other loved ones and to see what is going on in the lives of those that I really care about. It's just a great way to keep the connection going. I really do value the technology. [P78, M, 45, Facebook]

I signed up with Facebook about 2 years ago. I started because my child's sport was keeping parents informed about sport related information. Now I mostly use it to stalk people. I hate that I look at it all the time. It feels like a time-suck. I'm looking at it and I don't even know

why I keep scrolling through items, but I do. I'm trying to limit it to just using it to post garage sale related items [...] [P85, F, 35, Facebook]

As follow-up questions, we asked participants who or what had brought them to use the technology in question (2), why they think they continue using it (3) and whether they had ever thought about quitting using this technology (4). The last question (4) was later used to create groups of participants questioning their technology use versus those that did not. A systematic qualitative analysis of the open questions was beyond the scope of this paper. However, all answers are available in the online materials.

Measures

All measures consisted of 7-point Likert-type scales, unless otherwise noted. The items were presented in randomized order for each measure.

Construct validity was examined by exploring the relationship of the UMI with several established measures from SDT and user experience research. In SDT, the positive effects of need satisfaction on well-being are thought to be mediated by motivation [54]. Need satisfaction and vitality were expected to be in general more positively related to the self-determined types of regulation. The same general pattern was also expected for usability and likelihood to recommend. Satisfaction with life was included as a very global measure of well-being that is distant from technology use. Thus, it should ideally be not or only weakly related to the other measures.

Need Satisfaction

Need satisfaction is an essential aspect of positive user experiences with interactive technology [18, 31]. Satisfaction of the needs for autonomy (Cronbach's $\alpha = .83$), competence ($\alpha = .76$), and relatedness ($\alpha = .91$) were measured with three items each, taken from Sheldon's need satisfaction scale [49].

The introductory question was adapted to reflect the use of technology: *How do you feel when you use [X]?*. Perceived need satisfaction for autonomy ($M = 5.1$, $SD = 1.4$), competence ($M = 4.14$, $SD = 1.59$), and relatedness ($M = 4.36$, $SD = 1.87$) were around the middle of the scale. An overall need satisfaction score ($\alpha = .8$) aggregated over all three needs was calculated with an average of $M = 4.53$, $SD = 1.16$.

Vitality

To gain an understanding of how different motivations affect well-being, state level vitality ($\alpha = .92$) was measured with seven items developed by [44]. Item wording was slightly adapted to include the technology (e.g., *When I use [X], I feel alive and vital.*). Descriptive statistics showed that participants tended to answer this scale around the midpoint of the scale ($M = 4.25$, $SD = 1.4$).

Satisfaction with life scale (SWLS)

To measure a construct that was not related to use of technology directly, but might be related to feelings of need satisfaction and vitality, general life-satisfaction was measured with the five items developed by Diener et al. [13]. Internal consistency was high ($\alpha = .91$) and agreement was moderate ($M = 4.62$, $SD = 1.49$).

Usability Metric for User Experience (UMUX)

The four items of the UMUX developed by Finstad [14] were employed to measure perceived usability of the reported technology. Internal consistency was acceptable with $\alpha = .69$. Overall, perceived usability was high ($M = 6.12$, $SD = 0.86$).

Likelihood to recommend (LTR)

LTR is a measure of engagement and satisfaction that is distinct but related to usability [46]. LTR was assessed with the question used to calculate the Net Promoter Score [40] on a scale from 0 to 10. Average LTR was above 8 ($M = 8.47$, $SD = 1.94$).

Results

Item analysis

Descriptive statistics of the UMI items were in the same range as in Study 1 (see Table 3). Average agreement to the statement was higher and rather left-skewed for the more self-determined types of regulation than for amotivation, external regulation and introjected regulation. Inter-item correlations were high within the factors, but low to moderate between the different factors (see also additional Tables in the online materials).

Confirmatory factor analysis (measurement model)

To test the multidimensional factor structure of the UMI, a confirmatory factor analysis (CFA) was conducted. All items were specified to load on their designated factor, and the loading of the first item was constrained to one. Multivariate normality was not given (Mardia tests: $\chi^2_s = 3888.9$, $p < .001$; $Z_k = 32.278$, $p < .001$), therefore we used a robust maximum likelihood estimation method with Huber-White standard errors and a Yuan-Bentler based scaled test statistic. Results of the CFA suggest that the proposed model fits the data well [$\chi^2_{120} = 237.53$, $p < .001$, $\chi^2/df = 1.98$, $CFI = .966$, $SRMR = .046$, $RMSEA = .046$, $PCLOSE = .771$]. The measurement model is depicted in Figure 2.

	#	<i>M</i>	<i>SD</i>	<i>S</i>	<i>K</i>	<i>pv</i>
Amotivation	1.	2.19	1.60	1.25	0.50	31.3
	2.	2.23	1.60	1.23	0.55	31.8
	3.	2.11	1.49	1.30	0.76	30.1
External regulation	1.	2.28	1.73	1.19	0.20	32.6
	2.	1.98	1.53	1.66	1.90	28.3
	3.	2.07	1.54	1.40	0.88	29.6
Introjected regulation	1.	2.59	1.89	0.97	-0.32	37.0
	2.	2.49	1.86	1.02	-0.26	35.5
	3.	2.00	1.61	1.61	1.47	28.6
Identified regulation	1.	4.89	1.66	-0.48	-0.40	69.9
	2.	5.33	1.62	-0.84	-0.09	76.2
	3.	5.06	1.77	-0.60	-0.65	72.3
Integrated regulation	1.	3.52	1.90	0.13	-1.14	50.2
	2.	3.38	1.88	0.29	-0.99	48.3
	3.	3.61	1.87	0.11	-1.03	51.6
Intrinsic motivation	1.	5.82	1.29	-1.13	0.91	83.1
	2.	5.79	1.22	-1.05	0.96	82.8
	3.	5.82	1.28	-1.22	1.41	83.1

Table 3. Descriptive statistics of all items including all participants ($N = 460$). S = Skewness. K = Kurtosis. pv = Difficulty index. Higher difficulty values indicate that people on average agree with this item, while lower values indicate the opposite.

A model with two factors (amotivation and extrinsic-intrinsic spectrum) and a model with 3 factors (amotivation, controlled regulation consisting of external and introjected regulations, and autonomous regulation consisting of identified and integrated regulations as well as intrinsic motivation) were tested. Results show that the fit for the alternative models were significantly worse than for the proposed model (refer to the online materials for detailed information on model comparison).

	<i>M</i>	<i>SD</i>	<i>M_{tr}</i>	ρ_C	α	AVE	MSV
AMO	2.18	1.42	1.77	.90	.90	.74	.36
EXT	2.11	1.36	1.71	.82	.81	.61	.22
INJ	2.36	1.58	1.92	.86	.85	.68	.27
IDE	5.10	1.44	5.21	.82	.81	.60	.36
INT	3.50	1.63	3.44	.84	.83	.63	.35
IMO	5.81	1.13	6.00	.88	.87	.71	.20

Table 4. Means, standard deviations and trimmed means (20%), Congeneric reliability (ρ_C), Cronbach's α , Average Variance Extracted (AVE) and Maximum Shared Variance (MSV) for the UMI in Study 2.

Reliability, convergent and discriminant validity

As seen in Table 4, congeneric reliability and internal consistency were high ($\rho_C > .7$, Cronbach's $\alpha > .8$), indicating high reliability. For all subscales Average Variance Extracted (AVE) was above the threshold of .5, suggesting high convergent validity. Maximum Shared Variance (MSV) values were lower than the corresponding AVE scores, indicating high discriminant validity.

Motivation and Related Measures

To investigate the relationship and construct validity of the six types of regulation with other constructs, we calculated the mean for each UMI subscale for each participant. Descriptive statistics of the six motivations are depicted in Table 4. Because most of the measures were not normally distributed, we calculated Pearson correlations with bootstrapping (1000

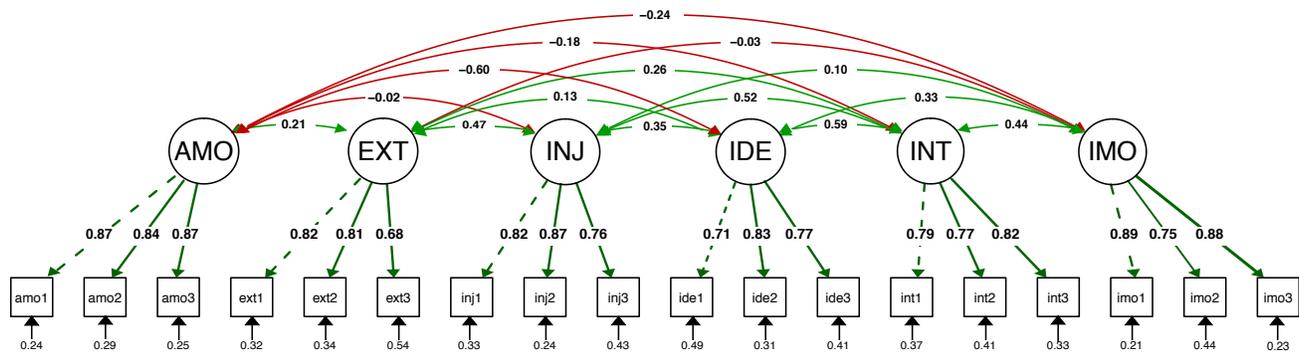


Figure 2. Measurement model of the UMI with standardized loadings. Dotted lines indicate loadings that were constrained to one. [$\chi^2_{120} = 237.53$, $p < .001$, $\chi^2/df = 1.98$, $CFI = .966$, $SRMR = .046$, $RMSEA = .046$, $PCLOSE = .771$]

iterations). Results are presented in Table 5. As posited by SDT, conceptually close regulations on the spectrum of motivation are more strongly correlated with each other than more distant regulations (termed 'simplex-like structure' in SDT literature). For the SDT related measures, need satisfaction and vitality were expected to be in general more strongly associated with the more self-determined regulations. The same pattern was also expected for usability and LTR since the more autonomous regulations are supposed to lead to the most positive outcomes. SWLS as a global measure of well-being was not expected to correlate substantially with technology related measures.

Need satisfaction

The most self-determined forms of regulation, integrated and intrinsic motivation, were positively associated with all need satisfaction measures. Relatedness was found to be positively associated with very self-determined regulations, but not with amotivation, identified or introjected regulations. Notably, relatedness was positively correlated with external regulation, which in SDT is not typically associated with positive outcomes on long-term motivation and well-being.

Vitality and Satisfaction with Life

Vitality showed positive correlations with all regulations except amotivation, which expectedly correlated negatively. In line with previous research within SDT (e.g., [58]), external regulation and vitality did not correlate significantly. Satisfaction with life was not directly related to any type of motivation, but was slightly positively correlated with the need for relatedness and feelings of vitality.

Usability

Amotivation and external regulation were negatively associated with perceived usability, whereas identified regulation and intrinsic motivation were positively correlated with usability. Neither introjected nor integrated regulation were significantly correlated with usability.

Likelihood to recommend

LTR was negatively correlated with amotivation, suggesting that users who do not know why they use a particular technology are less likely to recommend it to others. LTR was not

significantly related to external motivation, but positively associated with all remaining types of motivation. As expected, usability and LTR correlated positively.

Motivation and Questioning Technology Use

As a test of criterion validity, we investigated whether the UMI is able to detect differences between groups that we expected to differ in their motivation. The majority of the participants in Study 2 ($n = 297$) had never questioned their technology use, but 163 participants indicated that they had at some point questioned their use and thought about abandoning the technology, even though they were presently still using it. The two groups were compared with regards to their UMI ratings. Due to the non-equal group sizes and data featuring non-normal distribution, outliers and unequal variances, we applied robust Yuen-Welch tests to check for significant differences in trimmed means (as recommended by [1]). Results in Table 6 show that except for introjected and external regulation, all differences for the motivational regulations were statistically significant with effect sizes ranging from small (integrated regulation), over medium (intrinsic motivation, identified regulation) to large (amotivation). Participants who had questioned their technology use reported higher levels of amotivation, as well as lower levels of the more autonomous regulations and intrinsic motivation. This suggests that users who question their use have different regulations and might be more likely to abandon the technology in the future.

GENERAL DISCUSSION

The aim of the present work was to develop and validate a measure of motivation in the context of technology use based on self-determination theory. Our study results support the proposed factor structure of the UMI and show that it is a reliable and valid measure of users' motivation. The scale was found to have excellent psychometric properties measuring users' motivation across a wide range of technologies. Correlations of the UMI with related measures, most importantly need satisfaction, are in line with previous SDT research. We could show how different motivations relate to need satisfaction, usability and how they might affect consequences of technology use, such as well-being and likelihood to recommend. Moreover, the UMI was found to be sensitive to users who think about abandoning a technology. Lastly, by making all data and statistical scripts used in our analysis available

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
1.	Amotivation												
2.	External regulation	.20***											
3.	Introjected regulation	-.02	.37***										
4.	Identified regulation	-.51***	.09*	.30***									
5.	Integrated regulation	-.17***	.21***	.45***	.50***								
6.	Intrinsic motivation	-.23***	-.02	.11*	.29***	.40***							
7.	Autonomy	-.22***	-.05	.14**	.37***	.49***	.46***						
8.	Competence	-.29***	.05	.35***	.56***	.50***	.24***	.44***					
9.	Relatedness	.05	.31***	.04	.01	.29***	.22***	.33***	.07				
10.	Need satisfaction [†]	-.20***	.17***	.23***	.41***	.58***	.42***	.78***	.67***	.71***			
11.	Vitality	-.32***	.03	.32***	.46***	.58***	.47***	.51***	.63***	.28***	.64***		
12.	SWLS [‡]	-.05	.00	.02	-.02	.08	.03	.06	.08	.18***	.16***	.15**	
13.	Usability	-.43***	-.23***	-.06	.26***	.09	.33***	.32***	.14**	.03	.20***	.23***	.05
14.	LTR [§]	-.40***	.00	.14***	.46***	.34***	.44***	.43***	.38***	.10*	.40***	.43***	-.02
													.46***

Table 5. Pearson correlations with bootstrapping (1000 iterations) of the measures used in Study 2 (N = 460). Note. [†]Average of autonomy, competence, and relatedness. [‡]Satisfaction with life scale. [§]Likelihood to recommend. * p < .05; ** p < .01; *** p < .001.

	Use never questioned (n = 297)			Use questioned (n = 163)			Yuen-Welch test			
	M	SD	M _{tr}	M	SD	M _{tr}	t	df	p	ξ
AMO	1.74	1.054	1.38	2.98	1.649	2.79	-8.496	122.9	< .001	0.583
EXT	2.05	1.353	1.62	2.23	1.381	1.89	-1.831	178.6	.069	0.124
INJ	2.46	1.670	2.00	2.18	1.375	1.82	1.081	244.7	.281	0.085
IDE	5.38	1.342	5.54	4.58	1.473	4.58	5.849	180.2	< .001	0.404
INT	3.69	1.665	3.67	3.15	1.512	3.05	3.567	223.0	< .001	0.250
IMO	5.90	1.176	6.16	5.65	1.030	5.72	3.526	208.8	.001	0.257

Table 6. Comparison of participants who never questioned their use and participants who thought about quitting. M_{tr} = 20% trimmed means used for the Yuen-Welch test. ξ = Explanatory measure of effect size; interpretation: 0.10 small, 0.30 medium, 0.50 large.

for the CHI community [25], we hope our paper provides a helpful and transparent template for future scale development and validation endeavors in HCI. In the following, we discuss the theoretical implications of our findings. We also report limitations of the present work, outline future research directions, and provide practical instructions for applying the UMI.

Participants reported relatively high levels of the self-determined and autonomous regulations (i.e., identified, integrated, and intrinsic motivation) and low levels of nonself-determined and controlled regulations (i.e., amotivation, external and introjected motivation). This suggests that the motivation to use a specific technology was relatively self-determined on average, perhaps reflecting the presumably more self-determined use of technology in users’ spare time. However, about one third of participants had at some point thought about quitting the technology reported. Notably, they reported significantly lower levels of intrinsic motivation, integrated regulation, and identified regulation as well as higher levels of amotivation. This pattern is in line with the findings of Pelletier et al. [36], who showed that the more self-determined types of regulation were positively associated, whereas amotivation was consistently negatively associated with persistence. Moreover, while thinking about quitting a technology does not readily correspond to actually abandoning or even just intending to quit a technology, it is a step that may eventually lead to such a change in behavior. For instance, research on the motivation of high school students showed that less self-determined motivation correlated with higher levels of drop-out intentions, which was associated with actual drop out one year later [55].

Research in SDT also emphasized the importance of autonomy support as a predecessor of self-determined motivation and behavioral persistence [36, 55]. Correlations of the UMI with autonomy need satisfaction support this notion in the context of technology use, as autonomy was more strongly associated with self-determined regulations. In the SDT framework, these regulations are thought to link need satisfaction and affective, cognitive and behavioral consequences. With the UMI, researchers have a tool to investigate why people interact with a technology and possibly explain why a specific technology can have positive as well as negative effects on users’ experience, well-being and behavior. For instance, results of Study 2 show that users who reported higher levels of the self-determined regulations also indicated stronger feelings of vitality after technology use – a measure of well-being. In general, this pattern of positive effects of self-determined regulation was also found for usability and likelihood to recommend.

Interestingly, among the more self-determined types of regulation, integrated regulation was not related to usability, perhaps suggesting that when the use of a technology aligns with one’s values and core principles, usability might not be as important. In contrast, Mekler and Hornbæk [31] found that users who reported eudaimonically motivated experiences often mentioned instrumental qualities of a technology. However, note that Mekler and Hornbæk studied single experiences, whereas participants in the present work were asked to report on a frequently used technology and not a specific experience episode. Following previous SDT research, the UMI is a measure of general technology use, but evidence from other domains (e.g., academic achievement [17]) suggests that the UMI may also be applicable to single experience episodes. However, in spe-

cific situations motivated by one's personal values, usability might indeed be important (e.g., setting up a personal website). Taken together, it would be interesting to study how users' motivation shapes single experience episodes, and how experience episodes in turn influence motivation to use a technology. However, drawing from results in other domains (e.g., academic achievement [17]), situational motivational regulation can be expected to show a similar pattern as with general use. The UMI should be applied to examine motivation in single episodes to further test this notion.

SDT postulates that if people experience need satisfaction they can internalize an initially extrinsically motivated behavior (e.g., using a software because it is mandatory at work), shifting their motivational regulation from external towards integrated regulation over time. This means that over time people can feel effective in undertaking nonself-determined behaviors and they are more likely to personally endorse these actions. Autonomy supportive design [5], for instance, aims to design technologies that foster autonomy and self-determination over time. The UMI may thus be used to evaluate effects of different designs on motivation. Additionally, the UMI may possibly explain under what circumstances and for whom autonomy supportive technology is particularly effective. For instance, future research could study whether users with a controlled motivation or users with a more self-determined motivation benefit more from autonomy support. Although an in-depth investigation of these relationships was beyond the scope of the present work, our findings show that motivational regulations are associated differently with these constructs and provide a useful lens towards understanding user experience and outcomes of technology use.

With the UMI, researchers have a theory-based instrument to investigate the *why* of interaction. One advantage of approaching the why of interaction from the perspective of a well-researched theory is that one may draw from the large body of evidence on SDT to formulate and test hypotheses, and investigate this theory's applicability, limits, and predictive power for HCI [22]. The UMI may potentially help predict how likely users are to abandon a technology if given the opportunity. It may also serve to extend existing models of user experience that have already incorporated need satisfaction (e.g., [19, 22]), as different motivational patterns can explain why people have different experiences (and consequences) when using a particular technology. Specifically, the UMI may help to better understand the role of motivation in shaping need satisfaction, as well as how need satisfaction influences motivation to interact with a technology in the long term.

Limitations and Future Directions

The two studies reported here entail an initial thorough validation of the UMI. Although participants' age was distributed over a wide range, and a diverse set of technologies has been reported, the UMI needs to be further tested in-depth with users outside of Mechanical Turk and North America. The distinction between different types of regulation and their relation to well-being has been found to hold true in various languages and cultures [6, 15]. The structural validity of the UMI may be tested with other cultures to examine whether

this is also applicable for motivation in technology use. Most technologies reported by the participants were leisure-oriented. Thus, the structure of the UMI needs to be tested in the work-related technology use as well. However, existing evidence for differentiating motivational regulations in the work domain [15] is encouraging.

Additionally, the types of technology could be specified to different types of domains or even specific technologies. For instance, user's experience and behavior has been found to vary in the domains of fitness technologies [47], games [11], and Facebook [48], depending on whether they engaged with the technology to get recognition from others or because they had personally endorsed it. The UMI might allow for more nuanced insights into the motivational processes underlying users' experience and technology use.

In the present study, we examined only technology that users reported to use frequently, therefore limiting us to technology that has not (yet) been abandoned. In a next step, known-groups validity should be investigated, for instance by examining how users perceived abandoned technology, similar as in the validation of the User Burden Scale [51]. Since the UMI was found to be sensitive to users who think about abandoning a technology, it would be interesting to test whether the UMI is predictive for abandoning a technology. Finally, a promising avenue for future research is examining whether the UMI relates to behavioral intentions and, most importantly, actual behavior.

UMI Guidelines for Use

While the present studies employed a placeholder [X] for the 18 UMI items, this may be replaced with the technology under investigation (e.g., "Using Facebook is fun"). To reduce sequence effects, it is generally advisable to randomize the order of the items. We used a 7-point Likert-type agreement scale from 1 (strongly disagree) to 7 (strongly agree) and recommend using the same answering scale to ensure comparability. Researchers can calculate a score for each regulation separately by averaging the three items corresponding to the subscale.

CONCLUSION

We present the development and validation of a multidimensional measurement tool rooted in self-determination theory that helps to deepen our understanding of *why* users interact with a technology. The development and validation followed best practices and all data collected in the two studies together with the materials and analysis code is available online. The UMI has been tested with over 900 participants and shows promising psychometric properties, high reliability, convergent and discriminant validity. The UMI has implications for theory and practice and opens up opportunities for future research on motivation and user experience.

ACKNOWLEDGEMENTS

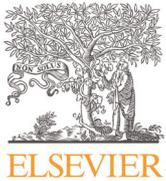
Special thanks to Kasper Hornbæk, Lena Aeschbach, Seamus Forde, Livia Müller, and Serge Petralito. This work has been approved by the Institutional Review Board of the Faculty of Psychology, University of Basel under the number D-003-17.

REFERENCES

1. Marjan Bakker and Jelte M. Wicherts. 2014. Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods* 19, 3 (2014), 409–427. DOI: <http://dx.doi.org/10.1037/met0000014>
2. Max V. Birk, Cheralyn Atkins, Jason T. Bowey, and Regan L. Mandryk. 2016. Fostering Intrinsic Motivation Through Avatar Identification in Digital Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2982–2995. DOI: <http://dx.doi.org/10.1145/2858036.2858062>
3. Marc R. Blais, Stéphane Sabourin, Colette Boucher, and Robert J. Vallerand. 1990. Toward a Motivational Model of Couple Happiness. *Journal of Personality and Social Psychology* 59, 5 (1990), 1021–1031. DOI: <http://dx.doi.org/10.1037/0022-3514.59.5.1021>
4. Ingwer Borg and Patrick J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Berlin: Springer Science & Business Media. DOI: <http://dx.doi.org/10.1007/0-387-28981-X>
5. Rafael A. Calvo, Dorian Peters, Daniel Johnson, and Yvonne Rogers. 2014. Autonomy in technology design. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*. ACM, 37–40. DOI: <http://dx.doi.org/10.1145/2559206.2560468>
6. Valery Chirkov, Richard M. Ryan, Youngmee Kim, and Ulas Kaplan. 2003. Differentiating autonomy from individualism and independence: A self-determination theory perspective on internalization of cultural orientations and well-being. *Journal of Personality and Social Psychology* 84, 1 (2003), 97–110. DOI: <http://dx.doi.org/10.1037/0022-3514.84.1.97>
7. Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (2016), 4–19. DOI: <http://dx.doi.org/10.1016/j.jesp.2015.07.006>
8. 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. DOI: http://dx.doi.org/10.1207/S15327965PLI1104_01
9. Edward L. Deci and Richard M. Ryan. 2002a. *Handbook of self-determination research*. Rochester, NY: University Rochester Press.
10. Edward L. Deci and Richard M. Ryan. 2002b. Overview of Self-Determination Theory. In *Handbook of self-determination research*, Edward L. Deci and Richard M. Ryan (Eds.). Rochester, NY: University Rochester Press, Chapter 1, 3–33.
11. Sebastian Deterding. 2016. Contextual Autonomy Support in Video Game Play : A Grounded Theory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3931–3943. DOI: <http://dx.doi.org/10.1145/2858036.2858395>
12. Robert F. DeVellis. 2017. *Scale Development: Theory and Applications*. Vol. 26. Thousand Oaks, CA: Sage publications.
13. Ed Diener, Robert A. Emmons, Randy J. Larsen, and Sharon Griffin. 1985. The Satisfaction With Life Scale. *Journal of Personality Assessment* 49, 1 (1985), 71–75. DOI: http://dx.doi.org/10.1207/s15327752jpa4901_13
14. Kraig Finstad. 2010. The Usability Metric for User Experience. *Interacting with Computers* 22, 5 (2010), 323–327. DOI: <http://dx.doi.org/10.1016/j.intcom.2010.04.004>
15. Marylène Gagné, Jacques Forest, Maarten Vansteenkiste, Laurence Crevier-Braud, Anja van den Broeck, Ann Kristin Aspeli, Jenny Bellerose, Charles Benabou, Emanuela Chemolli, Stefan Tomas Güntert, Hallgeir Halvari, Devani Laksmi Indiyastuti, Peter A. Johnson, Marianne Hauan Molstad, Mathias Naudin, Assane Ndao, Anja Hagen Olafsen, Patrice Roussel, Zheni Wang, and Cathrine Westbye. 2015. The Multidimensional Work Motivation Scale: Validation evidence in seven languages and nine countries. *European Journal of Work and Organizational Psychology* 24, 2 (2015), 178–196. DOI: <http://dx.doi.org/10.1080/1359432X.2013.877892>
16. Graham S. Gaine and Jennifer G. La Guardia. 2009. The unique contributions of motivations to maintain a relationship and motivations toward relational activities to relationship well-being. *Motivation and Emotion* 33, 2 (2009), 184–202. DOI: <http://dx.doi.org/10.1007/s11031-009-9120-x>
17. Frédéric Guay, Robert J. Vallerand, and Céline Blanchard. 2000. On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion* 24, 3 (2000), 175–213. DOI: <http://dx.doi.org/10.1023/A:1005614228250>
18. Marc Hassenzahl. 2010. Experience Design: Technology for All the Right Reasons. *Synthesis Lectures on Human-Centered Informatics* 3, 1 (2010), 1–95. DOI: <http://dx.doi.org/10.2200/S00261ED1V01Y201003HCI008>
19. Marc Hassenzahl, Sarah Diefenbach, and Anja Göritz. 2010. Needs, affect, and interactive products – Facets of user experience. *Interacting with Computers* 22, 5 (2010), 353–362. DOI: <http://dx.doi.org/10.1016/j.intcom.2010.04.002>
20. Marc Hassenzahl and Daniel Ullrich. 2007. To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers* 19, 4 (2007), 429–437. DOI: <http://dx.doi.org/10.1016/j.intcom.2007.05.001>

21. Kasper Hornbæk and Morten Hertzum. 2017. Technology Acceptance and User Experience: A Review of the Experiential Component in HCI. *ACM Transactions on Computer-Human Interaction* 24, 5, Article 33 (Oct. 2017), 30 pages. DOI: <http://dx.doi.org/10.1145/3127358>
22. Kasper Hornbæk and Antti Oulasvirta. 2017. What Is Interaction?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5040–5052. DOI: <http://dx.doi.org/10.1145/3025453.3025765>
23. Matt C. Howard. 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 (2016), 51–62. DOI: <http://dx.doi.org/10.1080/10447318.2015.1087664>
24. Matt C. Howard and Robert C. Melloy. 2016. Evaluating Item-Sort Task Methods: The Presentation of a New Statistical Significance Formula and Methodological Best Practices. *Journal of Business and Psychology* 31, 1 (2016), 173–186. DOI: <http://dx.doi.org/10.1007/s10869-015-9404-y>
25. Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanatham. 2017. Moving Transparent Statistics Forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17 (CHI EA '17)*. ACM, New York, NY, USA, 534–541. DOI: <http://dx.doi.org/10.1145/3027063.3027084>
26. Marc-André K. Lafrenière, Jérémie Verner-Filion, and Robert J. Vallerand. 2012. Development and validation of the Gaming Motivation Scale (GAMS). *Personality and Individual Differences* 53, 7 (2012), 827 – 831. DOI: <http://dx.doi.org/10.1016/j.paid.2012.06.013>
27. Effie Lai-Chong Law and Paul van Schaik. 2010. Modelling user experience – An agenda for research and practice. *Interacting with Computers* 22, 5 (2010), 313–322. DOI: <http://dx.doi.org/10.1016/j.intcom.2010.04.006>
28. Chris Lonsdale, Ken Hodge, and Elaine A. Rose. 2008. The behavioral regulation in sport questionnaire (BRSQ): instrument development and initial validity evidence. *Journal of sport & exercise psychology* 30, 3 (2008), 323–355. DOI: <http://dx.doi.org/10.1123/jsep.30.3.323>
29. Clifford Mallett, Masato Kawabata, Peter Newcombe, Andrés Otero-Forero, and Susan Jackson. 2007. Sport motivation scale-6 (SMS-6): A revised six-factor sport motivation scale. *Psychology of Sport and Exercise* 8, 5 (2007), 600–614. DOI: <http://dx.doi.org/10.1016/j.psychsport.2006.12.005>
30. David Markland and Vanessa Tobin. 2004. A Modification to the Behavioural Regulation in Exercise Questionnaire to Include an Assessment of Amotivation. *Journal of Sport and Exercise Psychology* 26, 2 (2004), 191–196. DOI: <http://dx.doi.org/10.1123/jsep.26.2.191>
31. Elisa D. Mekler and Kasper Hornbæk. 2016. Momentary Pleasure or Lasting Meaning?. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16 (CHI '16)*. ACM, New York, NY, USA, 4509–4520. DOI: <http://dx.doi.org/10.1145/2858036.2858225>
32. Helfried Moosbrugger and Augustin Kelava. 2007. *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer. DOI: <http://dx.doi.org/10.1007/978-3-642-20072-4>
33. Frederiki C. Moustaka, Symeon P. Vlachopoulos, Spyridoula Vazou, Maria Kaperoni, and David A. Markland. 2010. Initial validity evidence for the Behavioral Regulation in Exercise Questionnaire-2 among Greek exercise participants. *European Journal of Psychological Assessment* 26, 4 (2010), 269–276. DOI: <http://dx.doi.org/10.1027/1015-5759/a000036>
34. Elaine Mullan, David Markland, and David K. Ingledew. 1997. A graded conceptualisation of self-determination in the regulation of exercise behaviour: development of a measure using confirmatory factor analytic procedures. *Personality and Individual Differences* 23, 5 (1997), 745–752. DOI: [http://dx.doi.org/10.1016/S0191-8869\(97\)00107-4](http://dx.doi.org/10.1016/S0191-8869(97)00107-4)
35. Timo Partala and Alekski Kallinen. 2012. Understanding the most satisfying and unsatisfying user experiences: Emotions, psychological needs, and context. *Interacting with Computers* 24, 1 (2012), 25–34. DOI: <http://dx.doi.org/10.1016/j.intcom.2011.10.001>
36. Luc G. Pelletier, M. S. Fortier, Robert J. Vallerand, and N. M. Brière. 2001. Associations among perceived autonomy support, forms of self-regulations, and persistence: A prospective study. *Motivation and Emotion* 25, 4 (2001), 279–306. DOI: <http://dx.doi.org/10.1023/A:1014805132406>
37. Luc G. Pelletier, Meredith A. Rocchi, Robert J. Vallerand, Edward L. Deci, and Richard M. Ryan. 2013. Validation of the revised sport motivation scale (SMS-II). *Psychology of Sport and Exercise* 14, 3 (2013), 329–341. DOI: <http://dx.doi.org/10.1016/j.psychsport.2012.12.002>
38. Luc G. Pelletier, Kim M. Tuson, Isabelle Green-Demers, Kimberley Noels, and Ann M. Beaton. 1998. Why Are You Doing Things for the Environment? The Motivation Toward the Environment Scale (MTES)1. *Journal of Applied Social Psychology* 28, 5 (1998), 437–468. DOI: <http://dx.doi.org/10.1111/j.1559-1816.1998.tb01714.x>
39. Luc G. Pelletier, Kim M. Tuson, and Najwa K. Haddad. 1997. Client Motivation for Therapy Scale: a measure of intrinsic motivation, extrinsic motivation, and amotivation for therapy. *Journal of personality assessment* 68, 2 (1997), 414–35. DOI: http://dx.doi.org/10.1207/s15327752jpa6802_11
40. Frederick F. Reichheld. 2003. The one number you need to grow. *Harvard business review* 81, 12 (2003), 46–55.

41. Marco M. C. Rozendaal, David V. Keyson, and Huib de Ridder. 2007. Product Behavior and Appearance Effects on Experienced Engagement during Experiential and Goal-directed Tasks. In *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces*. ACM, 181–193. DOI : <http://dx.doi.org/10.1145/1314161.1314178>
42. Richard M. Ryan and James P. Connell. 1989. Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology* 57, 5 (1989), 749–761. DOI : <http://dx.doi.org/10.1037/0022-3514.57.5.749>
43. Richard M Ryan and Edward L Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25, 1 (2000), 54–67. DOI : <http://dx.doi.org/10.1006/ceps.1999.1020>
44. Richard M. Ryan and Christina Frederick. 1997. On Energy, Personality, and Health: Subjective Vitality as a Dynamic Reflection of Well-Being. *Journal of Personality* 65, 3 (1997), 529–565. DOI : <http://dx.doi.org/10.1111/j.1467-6494.1997.tb00326.x>
45. 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, 4 (01 Dec 2006), 344–360. DOI : <http://dx.doi.org/10.1007/s11031-006-9051-8>
46. Jeff Sauro and James R. Lewis. 2012. *Quantifying the User Experience: Practical Statistics for User Research*. Cambridge, MA: Morgan Kaufmann.
47. Hanna Schneider, Kilian Moser, Andreas Butz, and Florian Alt. 2016. Understanding the Mechanics of Persuasive System Design: A Mixed-Method Theory-driven Analysis of Freeletics. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 309–320. DOI : <http://dx.doi.org/10.1145/2858036.2858290>
48. Kennon M. Sheldon, Neetu Abad, and Christian Hinsch. 2011. A two-process view of Facebook use and relatedness need-satisfaction: Disconnection drives use, and connection rewards it. *Journal of Personality and Social Psychology* 100, 4 (2011), 766–775. DOI : <http://dx.doi.org/10.1037/a0022407>
49. Kennon M. Sheldon, Andrew J. Elliot, Youngmee Kim, and Tim Kasser. 2001. What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology* 80, 2 (2001), 325–339. DOI : <http://dx.doi.org/10.1037/0022-3514.80.2.325>
50. Kennon M. Sheldon, Richard M. Ryan, Edward L. Deci, and Tim Kasser. 2004. The Independent Effects of Goal Contents and Motives on Well-Being: It's Both What You Pursue and Why You Pursue It. *Personality and Social Psychology Bulletin* 30, 4 (2004), 475–486. DOI : <http://dx.doi.org/10.1177/0146167203261883>
51. Hyewon Suh, Nina Shahriaree, Eric B. Hekler, and Julie A. Kientz. 2016. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3988–3999. DOI : <http://dx.doi.org/10.1145/2858036.2858448>
52. Alexandre N Tuch, Paul Van Schaik, and Kasper Hornbæk. 2016. Leisure and Work, Good and Bad: The Role of Activity Domain and Valence in Modeling User Experience. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 6 (2016), 35.
53. Alexandre N. Tuch, Rune Trusell, and Kasper Hornbæk. 2013. Analyzing Users' Narratives to Understand Experience with Interactive Products. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2079–2088. DOI : <http://dx.doi.org/10.1145/2470654.2481285>
54. Robert J. Vallerand. 1997. Toward A Hierarchical Model of Intrinsic and Extrinsic Motivation. *Advances in Experimental Social Psychology*, Vol. 29. Academic Press, 271 – 360. DOI : [http://dx.doi.org/10.1016/S0065-2601\(08\)60019-2](http://dx.doi.org/10.1016/S0065-2601(08)60019-2)
55. Robert J. Vallerand, Michelle S. Fortier, and Frédéric Guay. 1997. Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology* 72, 5 (1997), 1161–1176. DOI : <http://dx.doi.org/10.1037/0022-3514.72.5.1161>
56. Paul van Schaik and Jonathan Ling. 2009. The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies* 67, 1 (2009), 79 – 89. DOI : <http://dx.doi.org/10.1016/j.ijhcs.2008.09.012>
57. Symeon P. Vlachopoulos, Ermioni S. Katartzi, Maria G. Kontou, Frederiki C. Moustaka, and Marios Goudas. 2011. The revised perceived locus of causality in physical education scale: Psychometric evaluation among youth. *Psychology of Sport and Exercise* 12, 6 (2011), 583–592. DOI : <http://dx.doi.org/10.1016/j.psychsport.2011.07.003>
58. Jia Wei Zhang, Ryan T. Howell, and Peter A. Caprariello. 2013. Buying Life Experiences for the "Right" Reasons: A Validation of the Motivations for Experiential Buying Scale. *Journal of Happiness Studies* 14, 3 (2013), 817–842. DOI : <http://dx.doi.org/10.1007/s10902-012-9357-z>



Measuring user rated language quality: Development and validation of the user interface Language Quality Survey (LQS)[☆]



Javier A. Bargas-Avila^{*}, Florian Brühlmann

Google/YouTube User Experience Research, Brandschenkestrasse 110, 8002 Zurich, Switzerland

ARTICLE INFO

Article history:

Received 5 January 2015
 Received in revised form
 24 August 2015
 Accepted 28 August 2015
 Communicated by E. Motta
 Available online 10 September 2015

Keywords:

User interface
 Language
 Text
 Translation
 Internationalization
 Localization
 L10n
 I18n

ABSTRACT

Written text plays a special role in user interfaces. Key information in interaction elements and content are mostly conveyed through text. The global context, where software has to run in multiple geographical and cultural regions, requires software developers to translate their interfaces into many different languages. This translation process is prone to errors – therefore the question of how language quality can be measured is important. This paper presents the development of a questionnaire to measure user interface language quality (LQS). After a first validation of the instrument with 843 participants, a final set of 10 items remained, which was tested again ($N = 690$). The survey showed a high internal consistency (Cronbach's α) of .82, acceptable discriminatory power coefficients (.34–.47), as well as a moderate average homogeneity of .36. The LQS also showed moderate correlation to UMUX, an established usability metric (convergent validity), and it successfully distinguished high and low language quality (discriminative validity). The application to three different products (YouTube, Google Analytics, Google AdWords) revealed similar key statistics, providing evidence that this survey is product-independent. Meanwhile, the survey has been translated and applied to more than 60 languages.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Key information in interaction elements and content within user interfaces are mostly conveyed through text. Graphical user interfaces have evolved substantially when compared to text-based user interfaces, but they still rely heavily on language to communicate with users. Therefore language plays a crucial role in Human–Computer Interaction. Single words can make the difference between failure or success.

The importance of language within a user interface (UI) becomes clear when text elements are removed. Fig. 1 shows three screenshots of the video-sharing site YouTube. The first (a) shows the original, the second (b) shows the website, but with all text elements removed, while on the third (c) all graphic elements are deleted. The illustration shows how the textless version is stripped of the most useful information: it is almost impossible to predict and choose which video to watch and navigation becomes impossible.

Text used in interfaces is highly dependent on cultural and regional aspects. For example, instructional text such as a tutorial could be worded informally for the US, but such an informal wording

might be very inappropriate in other cultures. Hence it is important to consider not only mere correctness of translation of text but also style and tone aspects in the specific cultural context. Beside translation of text, interface elements such as icons and pictures should also be considered in the process of localization. Worldwide, there are about 200 languages that are spoken by at least 3 million people (Lewis et al., 2013). Companies with worldwide reach need to localize their products to make sure they can be used by everyone. For instance, Google search currently supports more than 140, Facebook more than 60, and YouTube more than 60 languages.

Websites and user interfaces are generally developed in one source language and translated afterwards by professional linguists. The process of translation is prone to errors and might introduce a number of problems that are not present in the source user interface. For example, the word *auto* can be translated to French as *automatique* (automatic) or *automobile* (car), which obviously has a completely different meaning. Another problem arises from words that behave as a verb when placed in a button or as a noun if part of a label (Leiva and Alabau, 2014). For example, the word *access* can stand for “you have access” (as a label) or “you can request access” (as a button). This *word sense disambiguation problem* (Muntés Mulero and Paladini Adell, 2012) arises often in UI translations. Further, possible pitfalls are gender, prepositions without context (Muntés Mulero and Paladini Adell, 2012) or other characteristics of the source text that might influence the translation process (Dilts,

[☆]This paper has been recommended for acceptance by E. Motta.

^{*} Corresponding author.

E-mail addresses: javier.bargas@me.com (J.A. Bargas-Avila), florian.bruehlmann@gmail.com (F. Brühlmann).

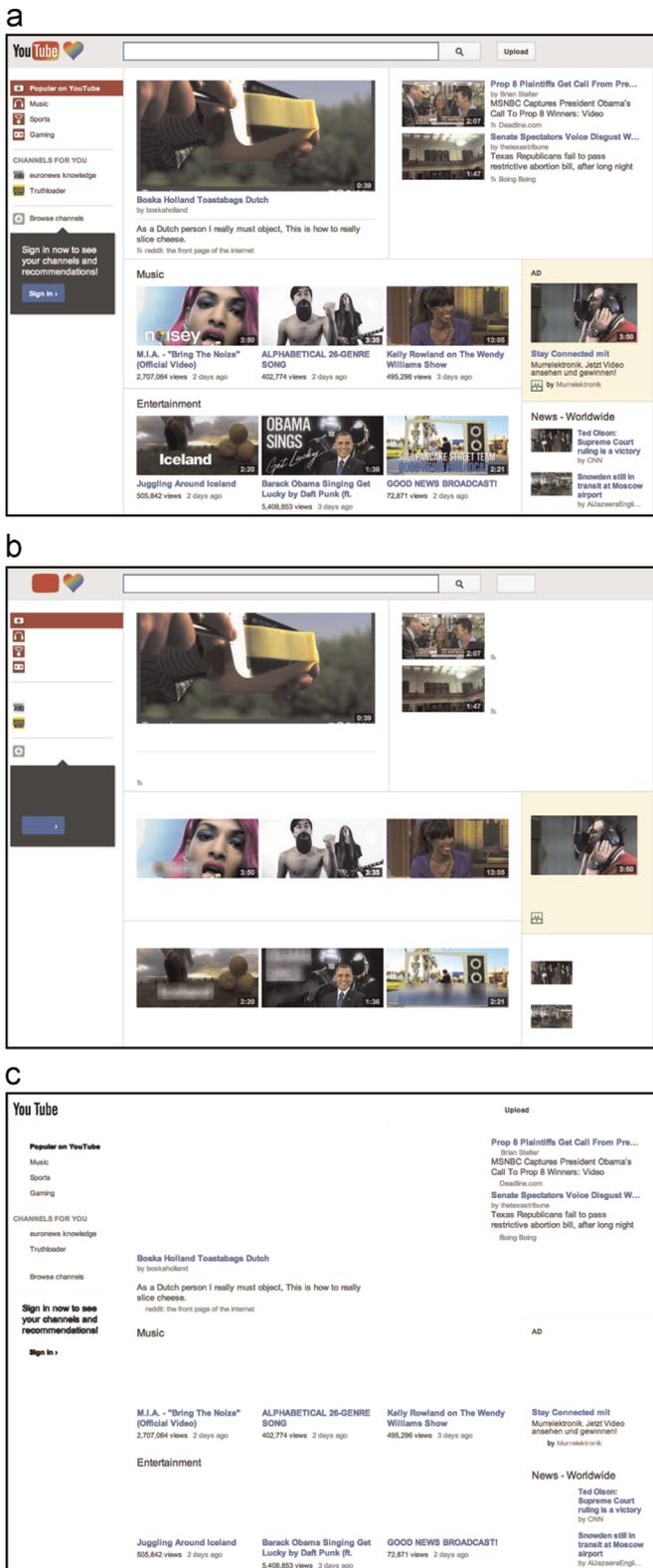


Fig. 1. Example of how UIs look when text or graphics are removed.

2001). Such mistranslations might not only negatively affect trustworthiness and brand perception, but also the acceptance of the website and its perceived usefulness (Sun, 2001).

As companies scale their products to multiple languages, the need for quality metrics increases: How can product managers learn more about the quality of a translation in an interface when they might not even speak the language themselves? In this paper,

a method is presented that delivers metrics about language quality by asking users to rate the language of the user interfaces in a survey.

2. Theoretical background

Schriver (1989) distinguishes three different classes of text quality evaluation: (1) text-focused, (2) expert-judgement-focused, and (3) reader-focused. These three classes express different levels on how explicit the feedback from the target audience is: "... text-focused methods (...) never use direct reader response; experts – through their experience – provide surrogate reader feedback; and reader-focused methods make explicit use of audience response." Schriver (1989, p. 241).

2.1. Text-focused evaluation

Text-focused methods operate by having a person or a computer examine a text and assess text quality by applying rules and guidelines that define what good text quality is. These methods include readability formulae (e.g., Fry, 1968; Kincaid et al., 1975) and user models (e.g., Blackmon et al., 2005; Chi et al., 2001) which can be applied by software that would allow automation of certain aspects of evaluation. Such automatized analysis is inexpensive and can spot certain obvious classes of error such as misspellings or provide general statistics about number of complex or passive sentences that could reduce readability. But in general, these provide little information about the overall performance of the text (whole-text level) or whether the text meets the needs of readers.

2.2. Expert-judgement-focused evaluation

Expert reviews involve a systematic screening of the text corpus by professional linguists. The major advantage of this method is that in-depth valuable feedback, which is based on expert knowledge, is produced. A drawback of this method can arise if evaluators are too close to the text or product that is examined, therefore making it harder to mentally take the users perspective when evaluating the language (Schriver, 1989). Also, this method is quite expensive to scale for products that are translated into many different languages.

2.3. Reader-focused evaluation

Schriver (1989) distinguishes two classes of reader feedback methods: (1) concurrent tests that evaluate the behaviors of readers in real-time, and (2) retrospective tests that are usually applied shortly after the reader has finished reading the text or after a certain time period. Concurrent methods include performance testing and thinking-aloud methods, while retrospective methods involve comprehension tests and surveys. Retrospective user testing is useful for revising existing text (Schriver, 1989).

Reader-focused methods have the advantage of giving information on global aspects of text quality and information about how the audience may respond to the text (Schriver, 1989). While retrospective methods such as surveys have disadvantages over concurrent methods (e.g., thinking-aloud or performance testing) because they rely on the use of memory, a survey during or after the interaction with a software might be a relatively reliable method to measure text quality. An empirical comparison of expert-focused and reader-focused methods of text evaluation showed that mutual agreement on problems in a text among experts is usually relatively low and contributed to a large set of false-alarms – problems that the readers did not report (Lentz and de Jong, 1997). This study also showed that experts experience difficulties with predicting the

problems that readers reported. The feedback of users is thus invaluable for judging the quality perception of text.

Schrivier (1989) argues that expert-judgement-focused evaluation should be used in combination with reader-focused evaluation methods to ensure the text comprehension of the target audience.

2.4. Background: How to focus on quality assurance resources?

In 2012, the YouTube internationalization team was in the following situation: anecdotal evidence suggested that some language versions of YouTube might benefit from improvement efforts. Past projects had shown that expert evaluations yielded good results and led to significant improvements of text quality. The problem with these evaluations was that they were time- and resource-consuming to conduct and analyze. The team did not have enough resources to conduct these reviews for all 60 languages and needed a reliable method to understand the state of each version.

User interface text is one among many aspects, such as date formats, color or icons and symbols, that need to be considered in the localization of a product. While there are guidelines for internationalization such as those proposed by del Galdo (1990), there are, to the authors knowledge, no validated scales available to specifically evaluate UI text quality.

Nielsen (1990) argues that a localized interface should be regarded as a *new* interface and therefore tested and analyzed accordingly. While task-based user testing of localized interfaces is important, users might not provide feedback about the language quality that goes beyond text errors encountered during a task. Also, usability testing with users for more than 60 language versions of an interface is very expensive and time consuming.

Based on this situation it was decided to apply a reader-focused method, and have YouTube users provide feedback on language quality through a survey. These data would then be used to determine which languages should be improved by expert evaluation efforts.

2.5. Six-subgroup quality scale

To the authors knowledge there is only one published scale that measures perceived text quality. The Six-Subgroup Quality Scale (SSQS) supports reviewers during the evaluation of an essay (Ransdell and Levy, 1996). It consists of six dimensions: (1) words: Choice and arrangement (readability), (2) technical quality: mechanics (tenses, grammar, spelling), (3) content of essay (engagement, egocentrism), (4) purpose/audience/tone (clear purpose, language and tone), (5) organization and development (elaboration, completeness, paragraphing), and (6) style (sentence structure, creativity).

While these quality criteria make sense for the evaluation of a multi-paragraph essay, not all of these aspects are relevant for user interface text. Many user interface text segments consist only of one word or a sentence. Applying, for instance, the categories “Content of essay”, “Organization and development” or “style” on user interface strings would yield little useful data.

Due to this situation, it was decided to develop and validate a survey to measure user interface language quality. The Language Quality Survey (short: LQS) aims to facilitate feedback for researchers and practitioners about the text quality of user interfaces and enable focused quality improvement efforts on problematic languages.

Note that this publication reports the development and validation of this survey. It does not report detailed results and findings regarding YouTube's language quality.

3. Development and first validation

3.1. Development of the LQS

3.1.1. Item-generation for the first version

In the first step, a group of professional linguists came together in a brainstorming session and discussed the core criteria of language quality. These linguists were experts in their field and involved in the process of user interface translation and validation. Only criteria that were unanimously accepted were included in the definition of language quality. The items of the questionnaire were then derived from the following formal definitions of language quality: friendliness, casualness, professionalism, naturalness, easy-to-understand, appropriateness, correctness and global satisfaction. The final set of items can be found in Table 1.

3.1.2. Scale

To reduce room for interpretation, cultural effects, and translation problems, it was decided to use a 5-point Likert-scale with fully labeled scale points. All scale labels can be found in Table 1.

3.1.3. Experimental procedure

In order to validate the LQS, it was implemented as an online survey and tested with English-speaking users from the US that were recruited on the platform YouTube with an in-product survey link. Participation was voluntary (opt-in) and no compensation was offered for taking part in the study. Users were asked to rate the text quality of the YouTube interface. All 10 items were presented in sequential order. At the end of the survey, users had the opportunity to provide open-text comments on the questionnaire. There were no major redesigns of YouTube during the time of measurement.

3.1.4. Sample and data cleaning

A total of 3588 participated in the survey. This sample was subject to a rigorous data cleaning procedure described here:

1. YouTube not only provides linguistic user-interface elements, but also large amounts of user-generated language. The survey instructions clarified that users should only think about user interface elements when answering the survey (“...would like you to think about the written language provided by YouTube in elements such as buttons, information dialogues, navigation or help text, not the text provided within video titles, descriptions, audio tracks or comments.”). To control whether users had read and followed this instruction, we asked them at the end of the survey: “Please tell us which of the following text elements came into your mind while rating the language quality of the YouTube interface”. With this procedure, a total of 2188 had to be removed because they indicated that they rated the language quality of user-generated content.
2. For this analysis we decided to include only native speakers. Therefore we asked participants to “Rate the level of your reading skills in English” (answers: Basic, Moderate, Fluent, Native). A total of 397 participants who did not choose “Native” were excluded.
3. Another important factor was whether users interact with the user interface often enough to make an accurate judgement of its language quality. Accordingly, a total of 15 participants were excluded because they indicated using YouTube less than once a week.
4. Because we wanted to assess only the English version of YouTube, people who indicated using YouTube also in non-English languages were removed from analysis. This was the case for 135 participants.

Table 1
The first version of the LQS.

No.	Item	Scale
1	How friendly or unfriendly is the text used in the [product name] interface? By “friendly” we mean that the language used shows that [product name] respects and likes their users	Very unfriendly; rather unfriendly; neither unfriendly nor friendly; rather friendly; very friendly
2	How casual or formal is the text used in the [product name] interface? By “casual” we mean that the language used is relaxed, like friends speaking to each other. By “formal” we mean that the language is academic, similar to the text of an essay or a legal document.	Very formal; rather formal; neither formal nor casual; rather casual; very casual
3	How professional is the text used in the [product name] interface? By “professional” we mean that the language is well-written and shows that [product name] cares about quality	Not at all professional; slightly professional; moderately professional; very professional; extremely professional
4	How natural or unnatural is the text used in the [product name] interface? Natural here means that the language used represents the way people normally speak to each other	Very unnatural; rather unnatural; neither unnatural nor natural; rather natural; very natural
5	How easy or difficult to understand is the text used in the [product name] interface?	Very difficult to understand; rather difficult to understand; neither difficult nor easy to understand; rather easy to understand; very easy to understand
6	How appropriate or inappropriate do you consider the text in the [product name] interface?	Very inappropriate; rather inappropriate; neither inappropriate nor appropriate; rather appropriate; very appropriate
7	How often do you encounter grammatical errors in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
8	How often do you encounter typos/spelling errors in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
9	How often do you encounter untranslated words that are not in English in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
10	How satisfied or dissatisfied are you with the quality of language in the [product name] interface when using English?	very dissatisfied; rather dissatisfied; neither dissatisfied nor satisfied; rather satisfied; very satisfied

Note: for this study, [product name] was replaced with “YouTube”.

5. Another 13 participants were removed because they could be identified as spam or left more than half of the items unanswered.

3.2. Results

The remaining sample consisted of $n = 843$ responses. The majority were male (73.5% male; 19.6% female; 6.9% did not indicate their sex) and 55.4% were between 18 and 29 years old. The gender distribution appeared to be skewed towards the male population. A comparison to the overall YouTube gender distribution was not possible, because there are no exact numbers (a significant amount of YouTube users do not provide their gender or age). The sample's demographic characteristics can be found in Table 2.

Table 3 offers an overview of all missing values for each item. To prevent further sample size reduction with listwise and pairwise deletion, the Expectation-Maximization Algorithm (EM) was used to replace missing values. EM is a valid and reliable method to replace missing values. It is generally preferred over listwise and pairwise deletion (Allison, 2002; Schafer and Graham, 2002) and is often used in survey validation research (Bargas-Avila et al., 2009, 2010).

Table 4 shows the statistics for the first validation. The distribution skewed negatively towards the higher end of the scale, therefore data were log-transformed for further analysis. Transformation is a widely used method to ensure normal distribution of data (Tabachnick and Fidell, 1996). The difficulty indices ranged between .69 and .86, which means that participants tended to answer the items positively.

According to Fisseni (2004) it is advisable to calculate the discriminatory power with a product-moment correlation of the item score with the test score for interval-scaled item responses. If the items of a scale have moderate to high positive corrected item-total correlations, one can expect that the items measure a similar construct as the total score of a questionnaire (Moosbrugger and Kelava, 2007). This means that in case of high discriminatory power, the respondents score for this item reflects the sum score of all other items for this particular respondent. The discriminatory power and

Table 2
Demographics of participants in the first validation.

Sex	N	%	Age	N	%
Female	165	19.6	17 or younger	126	14.9
Male	620	73.5	18–29	467	55.4
Not indicated	58	6.9	30–39	73	8.7
Total	843	100	40–49	37	4.4
			50–59	23	2.7
			60 or older	11	1.3
			Not indicated	106	12.6
			Total	843	100

Table 3
Missing values for each item.

Item	1	2	3	4	5	6	7	8	9	10
N	842	841	842	833	836	835	840	831	827	839
Missing	1	2	1	10	7	8	3	12	16	4
In %	.1	.2	.1	1.2	.8	.9	.4	1.4	1.9	.5

Cronbach's α for each item are listed in Table 5. The discriminatory coefficients ranged between .15 and .59 with a mean of .45 ($SD = .132$). Three items showed a coefficient below .50 (items 1, 2, 5 and 9). According to Borg and Groenen (2005) the lowest acceptable discriminatory power is .30. Item 2 showed a coefficient of .15. The rest of the items were in an acceptable to good range.

Homogeneity examines whether all items of the LQS measure the same construct (“language quality”) and whether there are items that overlap (measure similar aspects of the construct). We calculated this by averaging the inter-item correlations for each item (Briggs and Cheek, 1986) similar to the study by Bargas-Avila et al. (2009). The intercorrelation matrix (see Table 6) depicts this aspect with significant correlations for all items ($p < .01$) except for item 2 which showed non-significant correlations with items 6, 7, 8, 9 and

Table 4
Statistics, first validation (untransformed).

Item	M	SD	S	K	p_v
1	3.74	1.119	-.827	.135	.685
2	3.31	.981	-.258	-.406	.578
3	3.65	.944	-.551	-.021	.664
4	3.72	.983	-.702	.180	.682
5	4.24	.879	-1.263	1.617	.813
6	4.20	.921	-1.147	1.149	.803
7	4.40	.860	-1.854	4.061	.851
8	4.44	.867	-2.027	4.613	.864
9	4.28	.966	-1.310	1.175	.824
10	4.21	.929	-1.304	1.750	.805

Note: N=843; missing values=EM; $SE_S = .084$; $SE_K = .168$.
S=Skewness; K=Kurtosis; p_v =difficulty indices.

Table 5
Discriminatory power and Cronbach's α (first version).

Item	1	2	3	4	5	6	7	8	9
r_{it}	.408	.150	.506	.508	.482	.585	.523	.542	.365
α_{-i}	.751	.787	.733	.733	.738	.722	.732	.729	.755

Note: r_{it} =corrected item - total correlation; α_{-i} =Cronbach's α if item deleted; $\alpha_{item1-9}=.765$. N=843; missing values=EM.

Table 6
Intercorrelation matrix and homogeneity indices for item 1–10 (first version).

Item	1	2	3	4	5	6	7	8	9	10
1	1									
2	.278	1								
3	.348	.080*	1							
4	.319	.219	.321	1						
5	.264	.154	.354	.443	1					
6	.328	.051†	.441	.441	.477	1				
7	.164	-.038†	.346	.233	.207	.361	1			
8	.139	-.035†	.339	.247	.234	.385	.842	1		
9	.114	-.003†	.188	.187	.160	.271	.438	.496	1	
10	.275	.106†	.420	.370	.416	.477	.289	.311	.232	1
H	.248	.090	.315	.309	.301	.359	.316	.329	.231	.322

* Note: $p < .05$

† n.s.: unmarked correlations are significant ($p < .01$).

10 as well as a significant correlation with item 3 on a higher α -level ($p < .05$). The global item 10 showed moderate correlations in a range of .11–.48, with all items, with item 2 showing the lowest correlation (.11). The average homogeneity index for the scale was at .28 and the homogeneity indices for each item ranged from .09 to .36 with the lowest value for item 2 (.09). A possible explanation for the relatively moderate indices could be the complexity of the measured construct “language quality”, which is composed of many different aspects of language.

Cronbach's α for the LQS was moderate with .765, suggesting an acceptable reliability for the first version of this questionnaire. Item 10 was not included in the reliability analysis because it reflects a user's global evaluation of the language quality and could artificially inflate Cronbach's α . Table 5 shows that the internal consistency could be improved if item 2 is excluded.

3.3. Discussion of the first version of the LQS

The first validation of the LQS shows promising results. It also becomes clear that item 2 needs to be modified or deleted.

3.3.1. Scale

There is a tendency to use the LQS in the upper part of the five-point scale. This is not surprising, as YouTube is created and translated by professional linguists and therefore it can be expected that the language quality is rather good. Also this is in line with other research on satisfaction surveys which shows that these items are commonly answered in the upper part of the scales (Bargas-Avila et al., 2009).

3.3.2. Items

Item 2 showed insufficient statistical values in terms of low correlation with other items and unsatisfactory homogeneity index. The reliability of the questionnaire increases after deletion of this item. A closer analysis revealed that the wording “How casual or formal is the text used in the YouTube interface?” combined two aspects that are difficult to interpret. Casualness and formality are highly subjective aspects and might be perceived and judged very differently by different users. The low discriminatory power points at this problem, therefore item 2 was deleted.

The analysis of the open-ended question at the end of the questionnaire also revealed that some users reported encountering text that did not make sense in their opinion. This aspect was not yet covered with the LQS items. Hence, a new item was introduced for the next iteration, which would allow measuring the occurrence of nonsensical text (“How often do you encounter text that does not make sense?”).

4. Second validation

In the revised LQS the item “How casual or formal is the text used in the YouTube interface?” was removed and a new item, “How often do you encounter text that does not make sense?” was added (see Table 7 for a list of all items).

4.1. Experimental procedure

In order to validate the second version of the LQS, it was again implemented and tested in the same way the first version was validated.

4.1.1. Sample and data cleaning

A total of 3327 participants completed the survey. The same data cleaning as in the first study was applied. This way, 2161 participants had to be removed because they indicated that they rated the language quality of user-generated content. From the remaining sample, 333 were non-native English speakers, 7 did not use YouTube at least once a week, 95 used YouTube also in non-English languages, and 41 were removed because they could be identified as spam, left more than half of the items unanswered or answered all questions with the same value.

4.1.2. Results

The remaining sample consisted of $n = 690$ responses. As with the first study, the majority of the participants were male (75.9 % male; 17.8 % female; 6.2 % did not indicate their sex) and 59.6 % were between 18 and 29 years old (see Table 8).

Table 9 provides an overview of all missing values for each item. As described before, the Expectation-Maximization Algorithm (EM) was used to replace the missing values.

Table 10 shows the statistics for the second validation. As with the first version, the distribution of the item values skewed negatively towards the higher end of the scale, therefore data were log-transformed for further analysis. The difficulty indices ranged between .65 and .85, which reflects the participants' tendency to answer the items positively.

Table 7
The second version of the LQS.

No.	Item	Scale
1	How friendly or unfriendly is the text used in the [product name] interface? By “friendly” we mean that the language used shows that [product name] respects and likes their users	Very unfriendly; rather unfriendly; neither unfriendly nor friendly; rather friendly; very friendly
2	How professional is the text used in the [product name] interface? By “professional” we mean that the language is well-written and shows that [product name] cares about quality	Not at all professional; slightly professional; moderately professional; very professional; extremely professional
3	How natural or unnatural is the text used in the [product name] interface? Natural here means that the language used represents the way people normally speak to each other	Very unnatural; rather unnatural; neither unnatural nor natural; rather natural; very natural
4	How easy or difficult to understand is the text used in the [product name] interface?	Very difficult to understand; rather difficult to understand; neither difficult nor easy to understand; rather easy to understand; very easy to understand
5	How appropriate or inappropriate do you consider the text in the [product name] interface?	Very inappropriate; rather inappropriate; neither inappropriate nor appropriate; rather appropriate; very appropriate
6	How often do you encounter grammatical errors in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
7	How often do you encounter typos/spelling errors in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
8	How often do you encounter text that does not make sense in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
9	How often do you encounter untranslated words that are not in English in the text used in the [product name] interface?	Always; often; sometimes; rarely; never
10	How satisfied or dissatisfied are you with the quality of language in the [product name] interface when using English?	Very dissatisfied; rather dissatisfied; neither dissatisfied nor satisfied; rather satisfied; very satisfied

Note: For this study, [product name] was replaced with “YouTube”. Item no. 8 (bold) was added for this second version of the LQS.

Table 8
Demographics of participants in the second validation.

Sex	N	%	Age	N	%
Female	123	17.8	17 or younger	123	17.8
Male	524	75.9	18–29	411	59.6
Not indicated	43	6.2	30–39	56	8.1
Total	690	100	40–49	29	4.2
			50–59	11	1.6
			60 or older	4	.6
			Not indicated	56	8.1
			Total	690	100

Table 9
Missing values for each item (second version).

Item	1	2	3	4	5	6	7	8	9	10
N	689	689	683	684	684	682	676	678	672	679
Missing	1	1	7	6	6	8	14	12	18	11
In %	.1	.1	1	.9	.9	1.2	2	1.7	2.6	1.6

Table 10
Statistics, second validation (untransformed).

Item	M	SD	S	K	p_v
1	3.60	1.041	–.577	–.049	.651
2	3.64	.873	–.429	.130	.661
3	3.65	.983	–.665	.142	.664
4	4.16	.869	–.931	.595	.791
5	4.11	.920	–.895	.563	.779
6	4.37	.846	–1.449	2.083	.846
7	4.39	.852	–1.546	2.336	.853
8	4.14	.946	–.989	.535	.790
9	4.23	.983	–1.140	.500	.813
10	4.06	.936	–.980	.907	.770

Note: $N=690$; missing values=EM; $SE_S=.093$; $SE_K=.186$.
 S =Skewness; K =Kurtosis; p_v =difficulty indices.

The discriminatory power and Cronbach's α for each item are listed in Table 11. The discriminatory coefficients ranged between .39 and .63 with a mean of .52 ($SD=.085$). Five items showed a

Table 11
Discriminatory power and Cronbach's α (second version).

Item	1	2	3	4	5	6	7	8	9
r_{it}	.389	.490	.457	.499	.571	.634	.619	.606	.464
α_{-i}	.820	.806	.810	.805	.796	.790	.791	.792	.809

Note: r_{it} =corrected item - total correlation; α_{-i} =Cronbach's α if item deleted; $\alpha_{item1-9}=.820$. $N=690$; missing values=EM.

coefficient below .50 (item 1, 2, 3, 4 and 9). All items showed satisfactory values.

To explore the homogeneity, the intercorrelation matrix (see Table 12) depicts all significant correlations ($p < .01$). The global item 10 correlated in a range from .29 to .46 with all items, showing low to moderate correlations. The average homogeneity index for the scale is .36 and the homogeneity indices for each item ranged from .26 to .41. Compared to the first version of the LQS these values show an increase in the intercorrelations for all items.

Cronbach's α for the LQS was high with .820, suggesting very good reliability for the second version of this questionnaire. In most cases, values for Cronbach's α above .70 are acceptable to good, values between .80 and .90 are very good and values above .90 might indicate item redundancy (DeVellis, 2012). Again, item 10 was excluded from the reliability analysis. Table 11 shows that the internal consistency cannot be improved with the exclusion of any of the items.

4.1.3. Exploratory factor analysis

In order to investigate the structure of the items, a principal component analysis was conducted. Again, global item 10 was excluded from the analysis. The solution revealed two factors with an eigenvalue greater than 1.00, explaining 58.2% of the total variance. The factors were rotated using the Oblimin method with Kaiser Normalization. Oblimin rotation was chosen because it is reasonable to expect that the emerging factors are correlated. Analysis showed that the emerging factors correlated with $r=.429$. The factor scores of both factors, calculated with regression method, correlated significantly with the global item 10 ($r_{1(LC)}=.486$;

Table 12
Intercorrelation matrix and homogeneity indices for item 1–10 (second version).

Item	1	2	3	4	5	6	7	8	9	10
1	1									
2	.402	1								
3	.303	.262	1							
4	.215	.303	.482	1						
5	.326	.445	.436	.497	1					
6	.234	.341	.214	.269	.326	1				
7	.226	.306	.199	.248	.319	.849	1			
8	.209	.282	.310	.392	.330	.577	.580	1		
9	.162	.217	.197	.195	.275	.472	.485	.470	1	
10	.294	.410	.391	.461	.456	.418	.397	.428	.388	1
H	.263	.330	.310	.340	.379	.411	.401	.398	.318	.405

Note: All correlations are significant ($p < .01$).

Table 13
Exploratory factor analysis.

	Factor 1: linguistic correctness	Factor 2: readability
Eigenvalues	3.803	1.440
Friendly (item 1)	.249	.601
Professional (item 2)	.374	.646
Natural (item 3)	.236	.738
Easy to understand (item 4)	.315	.737
Appropriate (item 5)	.382	.778
Grammatical errors (item 6)	.901	.372
Typos/spelling errors (item 7)	.907	.345
Text does not make sense (item 8)	.769	.460
Untranslated words (item 9)	.707	.287

Note: Extraction method: principal component analysis.
Rotation method: Oblimin with Kaiser normalization.

$r_{2(R)} = .564$; $p < .001$). The factor loadings for the extracted factors are shown in Table 13. An interpretation based on factor loadings suggests that the first factor describes the frequency of (in)consistencies in the language (Linguistic Correctness) and the second factor describes how natural and smooth to read the used language is (Readability).

In conclusion, the data show evidence that the LQS has a bi-dimensional structure, covering the factors “Linguistic Correctness” (items 6–9) and “Readability” (items 1–5). The items associated with the two factors can be treated as sub-scales of a global language quality. The scores of the subscales correlate significantly with the global item 10 ($p < .01$) with $r = .507$ for linguistic correctness and $r = .573$ for readability. The reliability of the subscales is on an acceptable to good level with $\alpha = .836$ for linguistic correctness and $\alpha = .740$ for readability.

5. Validity and generalization

5.1. Convergent validity

Convergent validity was examined by exploring the relationship of the LQS with an established measurement of usability. In a study with a final set of $n = 211$ native English speakers on YouTube (same data cleaning applied as described in prior sections), participants answered the Usability Metric for User Experience (UMUX), before filling out the LQS (second and final version as described in Section 4). UMUX (Finstad, 2010) is a reduced version of the SUS (Brooke, 1996), and contains four items measuring perceived effectiveness, efficiency, satisfaction and overall usability. Finstad

showed that UMUX is a reliable, valid and sensitive alternative to SUS if a shorter metric is needed.

The reliability metrics of LQS and UMUX were high (Cronbach’s Alpha $\alpha = .829$ and $\alpha = .813$). The correlation of the overall LQS score with the convergent construct “usability” was moderate ($r = .396$, $p < .01$, $N = 211$). The LQS subscale Readability correlated with UMUX on a moderate level ($r = .446$, $p < .01$, $N = 211$), substantially stronger than the subscale Linguistic Correctness ($r = .157$, $p < .05$, $N = 211$).

Conceptually a moderate correlation between two related (but not identical) constructs is to be expected. A very low correlation would hint at the fact that language quality and usability are not correlated or that the LQS does not measure the targeted construct. A very high correlation would mean that both constructs overlap strongly and would question the necessity of a separate survey. In the case of LQS, the moderate correlations are evidence that it measures a construct that relates to usability, but is different enough to warrant a separate survey.

A possible explanation for the different correlation strengths could be that Readability contains aspects of language that are more directly related to usability. For instance ease of understanding or naturalness of the UI text might have a direct impact on product usability. In contrast, Linguistic Correctness, which describes aspects like typos or grammar errors, seems to impact ease of use less strongly.

These data provide evidence for convergent validity of the LQS. Language quality and usability are constructs that partly overlap, but are not the same. Language quality cannot be regarded as a stand alone aspect of a user interface – it clearly correlates with usability ratings, though on moderate levels.

5.2. Discriminative validity

To further examine the validity of the LQS, discriminative validity was examined. During data cleaning (see Section 4.1.1), all participants who indicated to have rated user-generated content (video titles, video descriptions or audio tracks) were removed from the analysis. To calculate discriminative validity, these data were used. It is reasonable to assume, that user generated language (UGL) will be of less quality than the expert-generated language of the YouTube user interface.

A sample of 430 participants who rated only UGL was identified. The average score (items 1–9) of this group is 3.36 ($SD = .756$). These levels are significantly lower than the score for the YouTube user interface ($\bar{x} = 4.03$, $SD = .593$, $N = 690$), $t(752.184) = 15.645$, $p < .001$, $d = .99$ (large effect).

This analysis provides further evidence, that the LQS is a valid tool to measure language quality. Participants who rated user-generated language provided significantly lower scores than users rating language that was created by experts.

5.3. Generalization to other languages

A key question of the LQS was: Would it scale to other languages and deliver valuable data? To answer this question, the LQS needed to be translated into other languages and new data had to be gathered.

To do this, the survey was translated for a selection of languages that show high YouTube usage. The survey was first translated by a professional linguist and then reviewed by a second one. Both translators received detailed instructions on aspects they should pay attention to. All parts that led to disagreement were discussed and resolved between the translators.

Table 14 shows a summary of the key statistics. The numbers show similar values for all languages. While the sample sizes vary, the number of missing data points is comparable and relatively

low for each language. Similar to the English version, item difficulties tend towards the higher end, which is probably due to the relatively high quality of the YouTube user interface language. The discriminatory power is satisfactory but some items were below the recommended value of .3 for Portuguese-Brazil and French. The homogeneity of the items in other languages is – similar to the English version – on the lower end and reflects the relatively complex construct of language quality. The values of Cronbach's α range between .755 and .849 which is an acceptable to good level.

Overall, the validation revealed that the translated versions of the LQS worked as expected and can be applied to measure user interface language quality.

5.4. Generalization to other products

To understand if the LQS can be generalized to other products than YouTube, we ran this study for two entirely different products: Google Analytics and Google AdWords. Analytics is a tool that allows website owners to track and understand their website traffic, AdWords is the platform that allows advertisers worldwide to buy, configure and track advertisement that is run on Google properties. If the LQS is product independent, key statistics should be similar, no matter if the surveys are answered by consumers (YouTube), website owners (Analytics) or advertisers (AdWords).

The item analysis for these two additional products revealed key statistic values that are close to the results for the YouTube Interface (see Table 15), providing evidence that the LQS can indeed be generalized to other products.

6. Case study: applying the LQS in the field

The main reason for developing the LQS was to discover problematic translations of the YouTube interface to allow focused quality improvement efforts. To do this, the LQS was translated to over 60 languages and data were gathered for all these versions of the YouTube interface. While the exact results for each language

are not the topic of this paper, a high level overview of the process and results are provided to practitioners:

- To understand quality of each UI version, we compared the results for the translated versions to the source language (here: English). We inspected first the global item, in combination with Linguistic Correctness and Readability. No further weighting was applied. Second, we inspected each item separately, to understand which notion of Linguistic Correctness or Readability showed worse (or better) values.
- The data revealed that about one third of the languages showed subpar language quality levels, when compared to the source language
- To understand the source of these problems and fix them, two actions were taken: (1) run a modified version of the LQS to gather qualitative feedback, and (2) conduct in-depth quality reviews with experts (as recommended by Schriver, 1989)
- The modified version of the LQS consisted of the identical survey, with one slight change. Every time a survey respondent selected the lower two end scale points, pointing to a problem in the language, a text box with the following question was surfaced: "Can you tell us what to improve? Any examples or links would help us understand what needs to be changed.". With this approach we aimed at generating more actionable qualitative knowledge on how to improve translations. The analysis of these comments provided linguists with valuable feedback of various kinds. For instance, users pointed to confusing terminology, untranslated words that were missed during translation, typographical or grammatical problems, words that were translated but are commonly used in English, or screenshots in help pages that were in English but needed to be localized. Some users also pointed to readability aspects such as sections with old fashioned or too formal tone as well as too informal translations, complex technical or legal wordings, unnatural translations or rather lengthy sections of text. In some languages users also pointed to text that was too small or criticized the readability of the font that was used. Experts did not always agree with the qualitative

Table 14
Statistics of the LQS in other languages.

Language	N	% mis (min)	% mis (max)	p_v (min)	p_v (max)	r_{it} (min)	r_{it} (max)	H (min)	H (max)	α_{1-9}
English (USA)	690	.1	2.6	.651	.840	.389	.634	.263	.411	.820
French (France)	308	1.9	5.2	.660	.870	.305	.593	.201	.377	.766
German (Germany)	1016	.6	2.5	.640	.850	.342	.554	.221	.329	.774
Italian (Italy)	896	.2	3.2	.690	.870	.329	.597	.217	.359	.793
Portuguese-BR (Brazil)	410	.7	5.4	.640	.850	.241	.592	.276	.340	.774
Russian (Russia)	358	.6	3.4	.730	.920	.406	.548	.253	.347	.781
Spanish (Spain)	333	.9	3.3	.610	.840	.451	.615	.274	.381	.825
Spanish LatAm (Mexico)	300	0	2.3	.640	.830	.429	.620	.310	.423	.844
Hebrew (Israel)	178	1.8	3.4	.669	.890	.379	.643	.260	.414	.828
Arabic (Saudi Arabia, Egypt, UAE, Morocco)	95	1.1	8.4	.580	.850	.394	.707	.270	.463	.849

Note: mis=missing values; p_v =item difficulty; r_{it} =discriminatory power; H =homogeneity; α =internal consistency.

Table 15
Generalization of LQS to other products.

Product	N	p_v (min)	p_v (max)	r_{it} (min)	r_{it} (max)	H (min)	H (max)	α_{1-9}
YouTube ^a	690	.651	.840	.389	.634	.263	.411	.820
Google Analytics ^a	902	.580	.880	.360	.616	.257	.431	.811
Google AdWords ^a	400	.670	.900	.368	.632	.249	.386	.809

Note: p_v =item difficulty; r_{it} =discriminatory power; H =homogeneity; α =internal consistency.

^a shown values are for LQS in English.

feedback from users. Many comments triggered fruitful conversations, of which not all led to changes.

- In parallel, in-depth expert reviews (the so-called “language find-its”) were organized. In these sessions, a group of experts for each language met and screened all of YouTube to discover aspects of the language that could be improved. All problems were gathered, discussed in the team, and concrete actions decided on how to fix them. By using the LQS data to select the target languages, it was possible to reduce the number of language find-its to about one third of the original estimation (if all languages had been screened).

In summary it can be said that the LQS proved a reliable, valid and useful tool to approach language quality evaluation and improvement.

7. Discussion

7.1. Summary and conclusions

There are three approaches to evaluate the quality of text (Schriver, 1989). (1) text-based evaluation methods such as automated readability scores can be easily calculated and are usually cost-effective, but their usefulness for improving language is rather superficial. (2) Expert-judgement based methods create in-depth actionable insights, but these approaches are limited due to the lack of an outside perspective, their difficulty in anticipating text problems on a user level and the high costs associated with them. (3) Reader-focused methods can be quite cost-efficient, provide user-centric perspectives, but generate few actionable insights on how to improve the language. Therefore, a combination of expert-judgement and reader-focused methods is promising.

This paper presents the development and validation of a reader-focused method: a survey enables companies to have their users rate the language quality provided in the user interface. Professional linguists agreed upon central aspects of language quality. Based on this, 10 items for the first version of the LQS were developed. This questionnaire was applied in an online survey in order to evaluate user interface text quality and to validate the questionnaire. The item analysis of the first version of the LQS revealed that one item did not satisfy statistical criteria and therefore was eliminated from the tool. After the first validation, qualitative user feedback suggested that the inclusion of an item to cover the occurrence of nonsensical text in the user interface would help users in the rating process. A new question was added to the questionnaire to measure this aspect.

The second version of the questionnaire showed good statistics. An exploratory factor analysis revealed that the questionnaire measures two factors: (1) more objective aspects such as typography, grammar and frequency of untranslated words that were summarized under the term *Linguistic Correctness* and (2) rather subjective aspects such as friendliness and appropriateness, named *Readability*.

Both validations of the LQS showed high Cronbach's α levels, which is a clear evidence of good internal consistency. The second validation indicates that Cronbach's α cannot be increased further by the exclusion of any item. For the second validation, the homogeneity indices have been increased to an acceptable level. Given the complexity of the construct “language quality”, heterogeneous items can be expected. Thus, the overall reliability and validity of the LQS are good. Good content validity can be assumed, as all items were developed and approved by a group of expert linguists, making it very likely that the most important aspects of language quality have been considered. Criterion-related validity is measured by the correlations with the global item, which also showed satisfactory

results. There is clear evidence for convergent validity, as shown by the correlation to UMUX, as well as discriminative validity, as shown by the analysis of user-generated vs expert-generated content. There is evidence that the validation of the LQS might be language-independent, because the analysis of other languages showed similar results. This survey can also be generalized to other products, because the application to Google Analytics and Google AdWords revealed similar survey validation statistics.

While it can be criticized that the questionnaire at hand measures language quality retrospectively, a concurrent reader-focused measure for the user interface language quality of a global website is not feasible and would be extremely expensive to accomplish. In general, the vast majority of questionnaires in the field of usability are applied post-use (Hornbaek, 2006).

In order to reach users worldwide, localization and translation are important. Even seemingly small differences such as having an Australian English version of a website as opposed to an international English version can make a difference for users: “And even in English-speaking Australia, users strongly preferred local sites to foreign sites. Although they could read both American and English-language European sites just fine, Aussie users felt that foreign sites were not as relevant to their needs.” (Nielsen, June 2011). While many other aspects of design such as color use, symbols and icons, as well as technical aspects such as date and time formats are important, a lot of the information is also conveyed through text.

del Galdo and Nielsen (1996) argue that there are three levels at which to tackle the problem of producing international user interfaces. The first level is the technical implementation of users' native language character set, notation and formats. This can be regarded as accomplished by most companies, according to del Galdo and Nielsen. The second level is producing a user interface and user information that are understandable and usable in the user's native language. The LQS aims to help reach this level by providing user-feedback about linguistic correctness and readability in order to assess and improve the text quality of a user interface. This is the foundation of the third level, proposed by del Galdo and Nielsen: the ability to produce systems that accommodate cultural characteristics of the users. This means that designs must address specific cultural models, such as the way people communicate or the way business is conducted in different countries.

The LQS allows practitioners to identify translations that need quality improvement which in turn allows the efficient allocation of resources to conduct expert-judgement based reviews. Also, the questionnaire can be applied at different stages of the product to measure the effect of changes. It is beneficial to combine the evaluation metrics with qualitative feedback. Allowing participants to provide reasons for their low rating on certain items has been proven to be useful for the derivation of actionable in-sights. The LQS has been extensively tested in the evaluation of the YouTube UI translation quality and helped to improve the language quality and ultimately the quality of the user experience.

Perceived language quality of translated user interfaces can have a significant impact on the perception of the overall quality and usability of a product. It is therefore important to assess and improve the quality of language used in applications. The LQS can be regarded as a small piece in the puzzle of understanding and improving language quality.

7.2. Limitations

There are several limitations of this study: (1) similar to most survey based approaches, participation was “opt-in”. This means that respondents could choose if they answer or not, which can lead to sampling biases. While this problem is present for almost all survey based approaches, it is important to keep in mind when

interpreting results. (2) In this publication, the LQS was applied only to browser based websites on desktop computers. Additional studies are needed to understand if it can be generalized to other applications, such as for instance mobile apps. (3) As stated before, there are several approaches to measure language quality. The LQS allows only a subjective user-based post-usage measurement and needs to be combined with other methods to deliver the full picture.

7.3. Future research

Future research could increase the validity of the survey by comparing post- to pre-revision results of the LQS. Practitioners and researchers might also benefit from a benchmark, which provides industry standards for good and bad LQS values. Another step could be to develop and validate a short version of the LQS that would allow measuring UI text quality in mobile context/applications.

Acknowledgements

We would like to thank the following people for their great help and support bringing this project to life: Devesh Kothari, Fredrik Lundh, Günther Noack, Keumhee Jeong, Matthew Glotzbach, Meike Schmidt, Olga Khroustaleva, Oliver Heckmann, Patricia Gómez Jurado, Svein Hermansen, and Wojtek Cyprys. Also, we would like to thank Sebastian Orsini for his help in survey validation methods, and Alexandre Tuch and Kasper Hornbæk for giving us feedback on early versions of the paper.

References

- Allison, P.D., 2002. Missing data: quantitative applications in the social sciences. *Br. J. Math. Stat. Psychol.* 55 (1), 193–196.
- Bargas-Avila, J.A., Lötscher, J., Orsini, S., Opwis, K., 2009. Intranet satisfaction questionnaire: development and validation of a questionnaire to measure user satisfaction with the intranet. *Comput. Hum. Behav.* 25 (November (6)), 1241–1250. <http://dx.doi.org/10.1016/j.chb.2009.05.014>.
- Bargas-Avila, J.A., Orsini, S., de Vito, M., Opwis, K., 2010. Zego: development and validation of a short questionnaire to measure user satisfaction with e-government portals. *Adv. Hum.-Comput. Interact.* (January 10), 6:1–6:10. <http://dx.doi.org/10.1155/2010/487163>.
- Blackmon, M.H., Kitajima, M., Polson, P.G., 2005. Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '05, ACM, New York, NY, USA, pp. 31–40. <http://dx.doi.org/10.1145/1054972.1054978>.
- Borg, I., Groenen, P.J., 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York.
- Briggs, S.R., Cheek, J.M., 1986. The role of factor analysis in the development and evaluation of personality scales. *J. Personal.* 54 (1), 106–148. <http://dx.doi.org/10.1111/j.1467-6494.1986.tb00391.x>.
- Brooke, J., 1996. Sus-a quick and dirty usability scale. In: *Usability Evaluation in Industry*, vol. 189, Taylor & Francis, London, p. 194.
- Chi, E.H., Pirolli, P., Chen, K., Pitkow, J., 2001. Using information scent to model user information needs and actions and the web. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '01, ACM, New York, NY, USA, pp. 490–497. <http://dx.doi.org/10.1145/365024.365325>.
- Del Galdo, E., 1990. Internationalization and translation: some guidelines for the design of human-computer interfaces. In: Nielsen, J. (Ed.), *Designing User Interfaces for International Use*, Elsevier Science Publishers Ltd., Essex, UK, pp. 1–10. (<http://dl.acm.org/citation.cfm?id=130347.132705>).
- Del Galdo, E.M., Nielsen, J., 1996. *International Users Interface*, John Wiley & Sons, Inc, New York.
- DeVellis, R.F., 2012. *Scale development: theory and applications*. Applied Social Research Methods Series, 3rd edition, vol. 26, Sage Publications, Newbury Park, CA.
- Dilts, D.W., 2001. Successfully crossing the language translation divide. In: Proceedings of the 19th Annual International Conference on Computer Documentation, SIGDOC '01, ACM, New York, NY, USA, pp. 73–77. (<http://dx.doi.org/10.1145/501516.501531>).
- Finstad, K., 2010. The usability metric for user experience. *Interact. Comput.* 22 (5), 323–327.
- Fisseni, H.J., 2004. *Lehrbuch der psychologischen Diagnostik: mit Hinweisen zur Intervention*, Hogrefe Verlag, Göttingen, Germany.
- Fry, E., 1968. A readability formula that saves time. *J. Read.* 11 (7), 513–578.
- Hornbæk, K., 2006. Current practice in measuring usability: challenges to usability studies and research. *Int. J. Hum.-Comput. Stud.* 64, 79–102.
- Kincaid, J.P., Fishburne Jr., R.P., Rogers, R.L., Chissom, B.S., 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report, DTIC Document.
- Leiva, L.A., Alabau, V., 2014. The impact of visual contextualization on ui localization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, ACM, New York, NY, USA, pp. 3739–3742. <http://dx.doi.org/10.1145/2556288.2556982>.
- Lentz, L., de Jong, M., 1997. The evaluation of text quality: expert-focused and reader-focused methods compared. *IEEE Trans. Prof. Commun.* 40 (September (3)), 224–234.
- Lewis, M.P., Simons, G.F., Fennig, C.D. (Eds.), *Ethnologue: Languages of the World*, 17th edition, SIL International, Dallas, 2013. URL (<http://www.ethnologue.com>).
- Moosbrugger, H., Kelava, A., 2007. *Testtheorie und Fragebogenkonstruktion*. Springer, Berlin, Germany.
- Muntés Mulero, V., Paladini Adell, P., España Bonet, C., Márquez Villodre, L., et al., 2012. Context-aware machine translation for software localization. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, EAMT 2012, EAMT, Trento, Italy, pp. 77–80.
- Nielsen, J., 1990. Usability testing of international interfaces. In: Nielsen, J. (Ed.), *Designing User Interfaces for International Use*, Elsevier Science Publishers Ltd., Essex, UK, pp. 39–44. (<http://dl.acm.org/citation.cfm?id=130347.132707>).
- Nielsen, J., June 2011. International Usability: Big Stuff the Same, Details Differ. (<http://www.ngroup.com/articles/international-usability-details-differ/>).
- Ransdell, S., Levy, C.M., 1996. Working memory constraints on writing quality and fluency. In: Levy, C. M., Ransdell, S. (Eds.), *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, Lawrence Erlbaum Associates, Hillsdale, NJ, Inc, pp. 93–105.
- Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. *Psychol. Methods* 7 (2), 147.
- Schrifer, K.A., 1989. Evaluating text quality: the continuum from text-focused to reader-focused methods. *IEEE Trans. Prof. Commun.* 32 (4), 238–255.
- Sun, H., 2001. Building a culturally-competent corporate web site: an exploratory study of cultural markers in multilingual web design. In: Proceedings of the 19th Annual International Conference on Computer Documentation, SIGDOC '01, ACM, New York, NY, USA, pp. 95–102. <http://dx.doi.org/10.1145/501516.501536>.
- Tabachnick, B., Fidell, L., 1996. *Using Multivariate Statistics*, 3rd edition. Harper Collins College Publishers, New York.

TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web

Florian Brühlmann, Serge Petralito, Denise C. Rieser, Lena F. Aeschbach, Klaus Opwis
Center for Cognitive Psychology and Methodology, Department of Psychology, University of Basel

Abstract

Trust is an essential factor in many social interactions involving uncertainty. In the context of online services and websites, the anonymity and lack of control make trust a vital element for successful e-commerce. Despite trust having received sustained attention, there is a need for validated questionnaires that can be readily applied in different contexts and with various products. We, therefore, report the development and validation of a semantic differential measuring users' trust on three dimensions. Compared to Likert-type scales, semantic differentials have advantages when it comes to measuring multidimensional constructs in various contexts. The TrustDiff measures users' perceptions of Benevolence, Integrity, and Competence of an online vendor with ten items. The scale was investigated in three independent studies with over 1000 participants and shows good structural validity, high reliability, and correlates expectedly with related scales. As a test of criterion validity, the TrustDiff showed significant differences on all subscales in a study involving a manipulated website.

Keywords: Trust, Semantic differential, Scale development, User experience, E-commerce

1. Introduction

Trust is an essential factor when acting under uncertainty and with the risk of negative consequences (Casaló et al., 2007). There are multiple definitions of trust in the literature, emanating from various academic fields (e.g., Driscoll, 1978; Moorman et al., 1993; Rotter, 1967). This renders a precise operationalization for measuring trust particularly challenging.

All definitions usually have two key components of trustworthiness in common: a willingness to be vulnerable, and a perception of the intentions of the other party (Lewicki and

8 Brinseld, 2012). The concept of trust on the web has important differences compared to trust
9 in offline contexts. Online trust is usually complicated by trust in the internet itself, and
10 the organization behind the technology. Additionally, trust is characterized by a lack of face
11 to face interaction, an asymmetry in the information available to each party, and concerns
12 about privacy (van der Werff et al., 2018). The question of whether trust in a web context
13 refers to the organization behind a website, individuals (who for example will select or deliver
14 your order), or to the internet technology itself (such as online payments) is still open for
15 debate (van der Werff et al., 2018). However, trust in a web context is usually built around
16 characteristics from e-commerce (Wang and Emurian, 2005). Accordingly, several question-
17 naires have been developed to measure trust (e.g., Bhattacharjee, 2002; Cho, 2006; Flavián
18 et al., 2006; Gefen, 2002; McKnight et al., 2002). One of the main issues of trust research
19 in web or e-commerce contexts is the lack of a common, validated, reliable, and versatile
20 measure (Kim and Peterson, 2017). We further identify several limitations of the above-
21 mentioned scales regarding applicability in research and practice. First, most questionnaires
22 incorporate Likert-type scales with domain-specific statements. For instance, the items de-
23 veloped by McKnight et al. (2002) are tailored to a specific website under examination (such
24 as “LegalAdvice.com is competent and effective in providing legal advice.”). To apply the
25 scales in a different context, it may be necessary to rephrase its items. However, rephrasing
26 the statements used in these questionnaires could result in a loss of reliability and validity.
27 Second, translating Likert-type statements into other languages can be a difficult and time-
28 consuming process, which may further affect validity. In the present work, therefore, we
29 describe the development and validation of TrustDiff; a semantic differential for measuring
30 trust on the web. This new measure displays several advantages over traditional Likert-type
31 scales when measuring complex and multidimensional constructs (Verhagen et al., 2015).
32 The results of three validation studies (total sample size $N = 1165$) indicate that TrustDiff
33 has excellent psychometric properties, measuring Benevolence, Integrity, and Competence
34 with high reliability. Furthermore, we demonstrate how these three subscales relate to an
35 existing Likert-type trust scale and the concepts of visual aesthetics and usability. Finally,
36 TrustDiff was found to be sensitive to the manipulation of trust-related features in an ex-

37 periment with a mock website. Taken together, TrustDiff represents a promising tool for
38 assessing trust in various domains of research and practice.

39 *1.1. Characteristics and dimensions of trust*

40 There are four characteristics of trust, which are generally observed and agreed upon
41 in the context of trust in e-commerce (Wang and Emurian, 2005). First, there must be
42 two specific parties in a trusting relationship – a trusting party (trustor, such as an online-
43 customer) and a party to be trusted (trustee, such as an online merchant). Second, trust
44 involves vulnerability, uncertainty, and risk for the trustor, while anonymity and unpre-
45 dictability are associated with the trustee. Third, trust leads to actions that are mostly
46 comprised of risk-taking behaviors, such as providing personal and financial information.
47 Finally, trust is subjective, and the level of trust considered sufficient for online transactions
48 is different for everyone. Moreover, people vary in their attitudes toward machines and tech-
49 nology (Wang and Emurian, 2005). Trust in e-commerce involves interpersonal trust, trust
50 in the organization representing a website, and trust in the underlying technologies (van der
51 Werff et al., 2018). In the Web Trust Model developed by McKnight et al. (2002), trusting
52 beliefs are at the core of what we consider the different dimensions of user trust. Although
53 there are multiple types of trusting beliefs found within the literature, three dimensions are
54 generally accepted (Bhattacharjee, 2002; Chen and Dhillon, 2003; Flavián et al., 2006; Gefen,
55 2002; Mayer et al., 1995; McKnight et al., 2002): *Benevolence*, *Integrity* and *Competence*.
56 Benevolence is related to the user belief that the other party is interested in their welfare,
57 is motivated by a search for a mutually beneficial relationship, and has no intention of op-
58 portunistic behavior. Integrity, sometimes referred as honesty Flavián et al. (2006), is the
59 belief that the other party is sincere and fulfills their promises. Finally, Competence implies
60 the other party has the resources and capabilities needed for the successful completion of
61 the transaction, and for the continuance of the relationship (Casaló et al., 2007).

62 *1.2. Existing questionnaires*

63 Various works have been directly or indirectly concerned with measuring trust (Bart
64 et al., 2005; Cho, 2006; Corbitt et al., 2003; Lee and Turban, 2001; Jarvenpaa et al., 1999;

65 McKnight et al., 2002; Pavlou and Gefen, 2004). However, from the practical and research
66 perspectives, there remains a need for a validated, brief, and easy-to-translate scale that mea-
67 sures trust and incorporates the three dimensions of Benevolence, Integrity, and Competence
68 (Kim and Peterson, 2017). The following problems with preexisting scales have been iden-
69 tified: First, not all of the existing scales inquire about trust directly; they often ask about
70 adjacent constructs such as Benevolence in Cho (2006), which in McKnight et al. (2002) is
71 merely a part of the model for trust. Second, existing measurement methods were created
72 to answer specific questions in certain contexts. An example of this is Lu et al. (2012), who
73 developed Likert-type questions for Customer-to-Customer (C2C) platforms, such as “Do
74 you agree that this C2C platform solves a security problem or stops a fraudulent behavior”.
75 Third, in their meta-analysis, Kim and Peterson (2017) described preexisting measurements
76 as “ambiguous” and stated that there is a necessity for a “well-developed scale to measure
77 online trust that is specifically tailored to the business-to-consumer e-commerce environ-
78 ment” (p. 52). Therefore, we decided to develop a semantic differential that addresses these
79 problems, and which also possesses certain advantages over Likert-scales.

80 *1.3. Advantages of semantic differentials*

81 Semantic differentials function by presenting respondents with a set of bipolar items
82 consisting of a pair of antonyms. This provides semantic differentials with specific advantages
83 over the more common Likert-style questionnaires. Respondents to Likert-type scales can
84 only indicate the extent to which they agree or disagree with a specific statement. Hence,
85 a respondent selecting “strongly disagree” does not necessarily imply they agree with the
86 opposite of the item (Chin et al., 2008). Conversely, the format of semantic differentials
87 enables respondents to express their opinion about a concept more fully; that is, ranging from
88 the negative polar to the positive polar. Another advantage is that semantic differentials can
89 reduce the acquiescence bias sometimes provoked by Likert-type scales (Friborg et al., 2006).
90 The acquiescence bias refers to a category of response biases indicating that respondents
91 have a tendency to agree with all items, or indicating a positive connotation (Friborg et al.,
92 2006). Additionally it has been demonstrated that semantic differentials outperform Likert-

93 based scaling regarding robustness (Hawkins et al., 1974), reliability (Wirtz and Lee, 2003),
94 and validity (Van Auken and Barry, 1995). Furthermore, semantic differentials function
95 effectively as a short-form scale format, which reduces survey completion time (Chin et al.,
96 2008). Finally, the literature suggests this format is appropriate for measuring complex and
97 multidimensional constructs (Verhagen et al., 2015).

98 **2. Development and validation strategy**

99 The development and validation followed the framework described by Verhagen et al.
100 (2015). In a first step, relevant literature and existing scales were reviewed to develop a
101 sample of bipolar scales reflecting the underlying concepts of Benevolence, Integrity, and
102 Competence. In the second step, linguistic and psychological bipolarity were established
103 through an extensive review by 18 experts. The scale anchors need to function as linguistic
104 and psychological antonyms in relation to the concept being measured. After these two steps,
105 a first study was conducted to reduce the item pool and establish the structural validity
106 (dimensionality) of the scale. We recruited 601 participants to conduct an exploratory
107 factor analysis, and to investigate correlations of the TrustDiff with related constructs.
108 This served as an initial test of discriminant and convergent validity. A second study with
109 312 participants was conducted to test the measurement model with a confirmatory factor
110 analysis, involving various types of interactive technology. The third study was set up as an
111 experiment with 252 participants, where trust-related elements of a website were actively
112 manipulated to test criterion validity.

113 **3. Item Pool Development and Review**

114 *3.1. Item pool*

115 The literature review identified several relevant trust questionnaires that were used as a
116 basis to develop an initial item. Key adjectives within sentences of existing questionnaires
117 were extracted (Bhattacharjee, 2002; Bart et al., 2005; Cho, 2006; Corbitt et al., 2003;
118 Flavián et al., 2006; Gefen, 2002; Gefen et al., 2003; Hong and Cho, 2011; Jian et al., 2000;

119 Koufaris and Hampton-Sosa, 2004; Lu et al., 2012; McCroskey and Teven, 1999; Pavlou
120 and Gefen, 2004; Rieser and Bernhard, 2016). Forty-three unique adjectives were identified,
121 several of them appeared multiple times in the literature. In a next step, possible antonyms
122 were selected with the help of dictionaries (www.merriam-webster.com, www.thesaurus.com,
123 www.leo.org) and near-duplicates were removed. This process resulted in 28 positive adjectives
124 with up to 3 different antonyms.

125 *3.2. Expert review*

126 An item-sort task as well as a test for linguistic and psychological bipolarity were per-
127 formed by an expert panel ($N = 18$) of trained psychologists and user experience researchers
128 using an online survey. Experts assigned each of the 28 adjectives to one of the three di-
129 mensions of trust. Adjectives assigned to the correct dimension by less than 13 participants
130 were excluded (Howard and Melloy, 2016). For each of the remaining adjectives, the best
131 fitting antonym with the highest agreement was chosen, resulting in an initial item pool of
132 20 items (refer to Table 1).

133 **4. Study 1**

134 The goal of Study 1 was to reduce the over-representative item pool by employing factor
135 analysis and test the convergent and discriminant validity of the scale.

136 *4.1. Method*

137 *Participants.* A total of 714 participants finished the online survey successfully. Responses
138 were excluded from the final data set according to the following criteria: First, if the response
139 time of the participant was under 150 seconds. Second, if a repeated response pattern (e.g.
140 crossing only the middle response option for a specific questionnaire) was detected. Third,
141 if by the end of the survey the participants themselves stated not to use the data for the
142 final data analysis. After data exclusion, responses from 601 participants (42% women, 58%
143 men, Mean age = 38 years, age range: 18 – 84) remained. Recruitment took place on
144 Amazon Mechanical Turk. For participation, the participants were reimbursed with \$0.60.

Table 1: Items of the trust questionnaire examined in Study 1.

	Item	M	SD	Mdn	S	K
Benevolence						
BEN1	ignoring – caring	4.49	1.241	4	−0.04	−0.51
BEN2	malicious – benevolent	4.49	1.253	4	−0.05	−0.28
BEN3	rude – cordial	5.08	1.158	5	−0.27	−0.24
BEN4	insensitive – sensitive	4.32	1.202	4	−0.03	0.18
BEN5	inconsiderate – empathic	4.52	1.221	4	−0.11	−0.03
Integrity						
INT1	dishonest – honest	4.82	1.356	5	−0.46	−0.13
INT2	insincere – sincere	4.75	1.304	5	−0.45	0.07
INT3	dishonorable – honorable	4.62	1.333	5	−0.22	−0.34
INT4	unbelievable – believable	5.08	1.376	5	−0.71	0.18
INT5	untruthful – truthful	4.93	1.364	5	−0.40	−0.36
INT6	fraudulent – credible	5.06	1.432	5	−0.58	−0.26
Competence						
COM1	clueless – knowledgeable	5.56	1.169	6	−0.91	1.04
COM2	incompetent – competent	5.51	1.211	6	−0.75	0.33
COM3	unskilled – skillful	5.39	1.178	5	−0.59	0.15
COM4	unqualified – proficient	5.39	1.193	6	−0.70	0.45
COM5	incapable – capable	5.55	1.196	6	−0.78	0.58
COM6	uninformed – informed	5.48	1.204	6	−0.65	0.26
COM7	inexperienced – experienced	5.60	1.221	6	−0.89	0.62
COM8	ineffective – effective	5.51	1.244	6	−0.88	0.66
COM9	inept – resourceful	5.43	1.225	6	−0.78	0.53

Note. M = Mean, SD = Standard deviation, Mdn = Median, S = Skew, K = Kurtosis. $N = 601$.

145 Only workers from Amazon Mechanical Turk living in the United States were eligible to
146 participate in the survey.

147 *Procedure and Materials.* Participants were asked to perform two tasks on one of two
148 randomly assigned websites. The first group received a link to an online shop (<http://www.crazysales.com.au>), where they were asked to find a product of their liking and to
149 inform their self about the return policy of the company. The second group was given a link
150 to a website (<http://www.sunshineloans.com.au>), which specializes in small loans (refer
151 to section 4.1). While using this website, the participants were asked to inform themselves
152 about loan costs and whether or not security is required to apply for a loan. These two
153 websites were chosen to assess trust in a realistic setting. Both websites were selected by
154 considering both, the website traffic and the website ranking (data from www.alexacom.com
155 and www.similarweb.com) in the United States, since the target audience of the survey was
156 inhabitants of the United States and the aim was to select relatively unknown websites to
157 prevent any biases from previous experience. Upon returning to the survey, participants
158 were asked to rate the website regarding trust (TrustDiff and a Likert-type Trust scale),
159 usability, and visual aesthetics. Finally, general demographic questions were presented.
160

161 *4.2. Measures*

162 All items from the below-mentioned questionnaires were presented in random order
163 within their own subsection of the survey. All measures consisted of 7-point Likert-type
164 scales ranging from 1 (strongly disagree) to 7 (strongly agree), unless otherwise noted.

165 *TrustDiff.* The 20 item-pairs of the initial Version of the TrustDiff were presented as a
166 semantic differential with seven steps between the antonym pairs. Seven steps were chosen
167 because it corresponds to the commonly used seven-point Likert scale and because it has
168 been successfully applied in other semantic differentials (e.g., Hassenzahl et al. (2003)).
169 Participants were instructed to rate the website owner (“Please rate the website owner on
170 the following dimensions”).

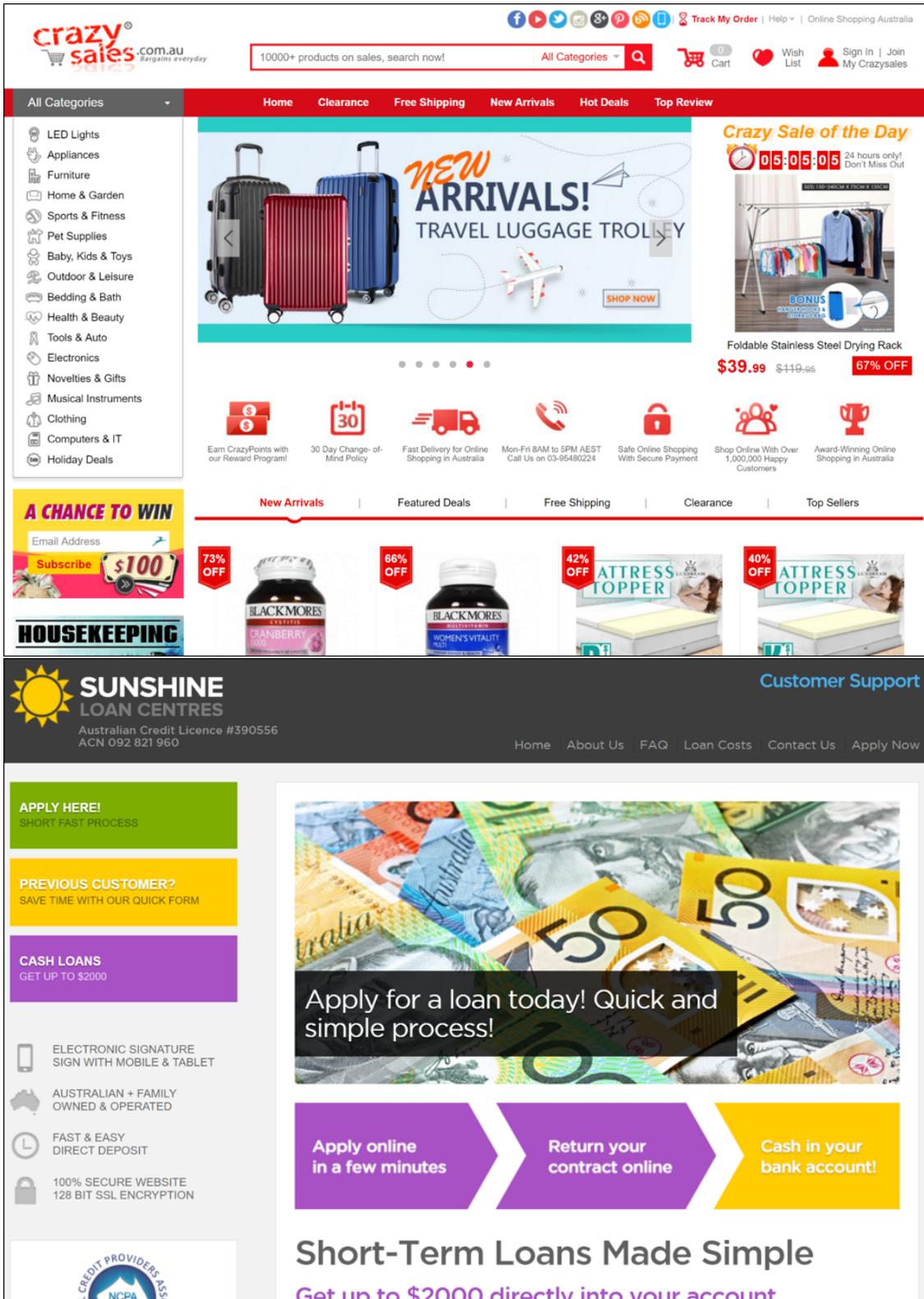


Figure 1: Screenshots from www.crazysales.com.au and www.sunshineloans.com.au

171 *Convergent trust scale.* To assess the convergent validity, the fifteen items of the trust ques-
172 tionnaire developed by Flavián et al. (2006) were included in the survey. Just like the
173 TrustDiff, this Likert-type scale measures trust with the three subscales Benevolence, In-
174 tegrity, and Competence. Slight modifications of the items' declarative statements were
175 carried out to better fit the measured website. The scale showed excellent internal consis-
176 tency: Benevolence ($\alpha = .90$), integrity ($\alpha = .90$), and competence ($\alpha = .90$).

177 *Visual aesthetics.* The discriminant validity of visual aesthetics was assessed, using eighteen
178 items of the VisAWI (Moshagen and Thielsch, 2010). To keep the analysis simple, all items
179 were average in an overall aesthetics score. Internal consistency was excellent for this scale
180 (Cronbach's $\alpha = .96$).

181 *Usability Metric for User Experience.* As an additional measure of discriminant validity
182 usability was measured using the four items of the Usability Metric for User Experience
183 (UMUX) (Finstad, 2010). Internal consistency of the scale was good (Cronbach's $\alpha = .87$).

184 4.3. Results

185 The full set of $N = 601$ was considered for the item analysis and exploratory factor
186 analysis. A two-samples Kolmogorov-Smirnov test was conducted to make sure that the
187 distributions in the data sets from each website did not differ significantly ($D = 0.090, p =$
188 $.169$). The item analysis and reduction process followed three steps. First, the distribution
189 statistics for each item were analyzed (see Table 1). Three items (COM1, COM7, COM8)
190 show a slight negative skew, suggesting a ceiling effect. For this reason and since competence
191 was measured using many items (9), they were excluded from further analysis.

192 Second, an exploratory factor analysis was conducted on the 17 remaining items with
193 oblique rotation, since factors were expected to be correlated. The Kaiser-Meyer-Olkin
194 measure verified the sampling adequacy for the analysis, $KMO = .97$ ('marvelous' according
195 to Hutcheson and Sofroniou (1999)), and all KMO values for individual items were greater
196 than $.95$, which is well above acceptable limit of $.5$ (Field, 2013). The Bartlett's Test
197 of sphericity, which tests the overall significance of all correlations within the correlation

198 matrix, was significant ($\chi^2(136) = 9533.923, p < .001$), suggesting that using an exploratory
199 factor analysis is appropriate. In an initial analysis of the eigenvalues only two factors
200 had eigenvalues over Kaiser's criterion of 1. However, the parallel analysis and the screen
201 plot suggested three factors which in combination explained 61% of the variance. The
202 exploratory factor analysis was performed using three factors, as this solution is in line
203 with the theoretical model of three subcomponents of trust. After the first exploratory
204 analysis, a total of three items (BEN3, INT2, and INT3) were eliminated because they did
205 not contribute to the factor structure and failed to meet the minimum criteria (Howard,
206 2016) of having a primary factor loading of .4 or above, and no cross-loading of .3 or above
207 (see Table 2).

208 A second exploratory factor analysis of the remaining 14 items, again with minres and
209 oblimin rotation, was conducted. The three factors explained 74% of the variance. All items
210 had primary loadings above .5 and load with their corresponding factor. The factor loadings
211 are presented in Table 3 and the correlations between the factors are presented in Table 4.
212 Finally, the reliability of each subscale was analyzed. Benevolence ($\alpha = .89$), integrity
213 ($\alpha = .95$), and competence ($\alpha = .93$) showed high internal consistency. No substantial
214 increase in Cronbach's alpha for any of the scales could have been achieved by eliminating
215 more items.

216 4.4. *Convergent and discriminant validity*

217 To assess convergent and discriminant validity the correlation of the TrustDiff and related
218 measures was explored. Table 5 depicts that the TrustDiff correlates strongly ($r = .68$) with
219 the trust questionnaire adapted from Flavián et al. (2006) indicating convergent validity.
220 The TrustDiff scale was found to correlate with visual aesthetics as well as usability ($r = .46$
221 and $r = .50$ respectively). Interestingly, the subscale Benevolence was less strongly related
222 to visual aesthetics and usability than the other subscales ($r = .33$ and $r = .34$ compared
223 to correlations in the range of .41 – .57).

Table 2: Rotated pattern matrix of the exploratory factor analysis in Study 1.

Item	Factor loadings			h2
	Benevolence	Integrity	Competence	
BEN1: ignoring – caring	.774	.079	.071	.767
BEN2: malicious – benevolent	.616	.179	.054	.629
BEN3: rude – cordial	.446	.070	.319	.530
BEN4: insensitive – sensitive	.848	–.018	–.005	.691
BEN5: inconsiderate – empathic	.860	.000	.016	.753
INT1: dishonest – honest	.143	.849	–.076	.830
INT2: insincere – sincere	.401	.508	.012	.741
INT3: dishonorable – honorable	.472	.430	.042	.764
INT4: unbelievable – believable	.086	.693	.082	.671
INT5: untruthful – truthful	–.035	.732	.160	.701
INT6: fraudulent – credible	–.035	.747	.205	.768
COM2: incompetent – competent	–.047	.126	.823	.791
COM3: unskilled – skillful	.099	–.084	.846	.707
COM4: unqualified – proficient	–.004	.067	.828	.763
COM5: incapable – capable	–.062	.114	.841	.793
COM6: uninformed – informed	–.027	.076	.832	.760
COM9: inept – resourceful	.125	–.145	.868	.699
Eigenvalues	1.98	0.73	10.46	
% of variance	18	17	26	

Note. Exploratory factor analysis with minres and oblimin.

Factor loadings above .3 are marked in bold. $N = 601$.

Three factors explain 61% of the total variance. $h2 = \text{Communality}$, $N = 601$.

Table 3: Results of the second exploratory factor analysis in Study 1.

Item	Factor loadings			h ²
	Benevolence	Integrity	Competence	
BEN1: ignoring – caring	.785	.081	.059	.779
BEN2: malicious – benevolent	.605	.174	.058	.611
BEN4: insensitive – sensitive	.825	.005	–.014	.675
BEN5: inconsiderate – empathic	.903	–.025	.009	.790
INT1: dishonest – honest	.143	.877	–.113	.834
INT4: unbelievable – believable	.074	.709	.060	.657
INT5: untruthful – truthful	–.011	.762	.121	.714
INT6: fraudulent – credible	–.030	.770	.173	.774
COM2: incompetent – competent	–.040	.121	.822	.793
COM3: unskilled – skillful	.087	–.065	.836	.700
COM4: unqualified – proficient	–.002	.065	.827	.762
COM5: incapable – capable	–.051	.097	.847	.795
COM6: uninformed – informed	–.014	.055	.841	.764
COM9: inept – resourceful	.114	–.142	.871	.693
Eigenvalues	0.70	1.80	8.62	
% of variance	18	18	38	
α	.90	.92	.95	

Note. Three factors explain 74% of the total variance.

Factor loadings above .3 are marked in bold. N = 601.

Table 4: Correlations between the factors extracted in Study 1.

Factor	Benevolence	Integrity	Competence
Benevolence	–		
Integrity	.76	–	
Competence	.52	.72	–

Note. N = 601.

Table 5: Descriptive statistics and Pearson correlations of measures in Study 1.

	M	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
TrustDiff												
1. Benevolence	4.45	1.08	–									
2. Integrity	4.97	1.24	.74	–								
3. Competence	5.46	1.07	.55	.72	–							
4. Total	4.96	1.00	.86	.94	.85	–						
Flavián et al. (2006)												
5. Benevolence	4.79	1.21	.60	.66	.54	.68	–					
6. Integrity	4.83	1.18	.54	.69	.54	.67	.86	–				
7. Competence	5.24	1.18	.36	.47	.57	.53	.75	.78	–			
8. Total	4.95	1.11	.54	.65	.59	.68	.93	.95	.91	–		
9. VisAWI	4.74	1.22	.33	.41	.47	.46	.46	.48	.49	.51	–	
10. UMUX	5.42	1.21	.34	.47	.53	.50	.48	.51	.55	.55	.75	–

Note. $N = 601$. All correlations are significant $p < .001$

224 4.5. Discussion

225 In Study 1, fourteen items measuring three related subcomponents of trust were identi-
226 fied. Analysis of correlations with related measures such as an existing rating scale offers a
227 first test of convergent validity. Comparatively low correlations of the TrustDiff with visual
228 aesthetics and usability indicate discriminant validity. The results of the second exploratory
229 factor analysis support a three dimensional measure with high reliability and good psycho-
230 metric properties. The measurement model of the TrustDiff was tested and refined in Study
231 2. The ability of the final TrustDiff to differentiate between two manipulated websites was
232 investigated in Study 3.

233 5. Study 2

234 5.1. Method

235 *Procedure and Measures.* As part of a larger study, participants were asked to name a single
236 interactive technology they use frequently. Participants indicated how often they had used
237 this particular technology over the last 14 days. The rest of the online survey focused on
238 this particular technology and the 14 items of the TrustDiff in Study 1 were included. As
239 in Study 1, the word pairs were presented in random order.

240 *Participants.* A total of 315 participants from the United States completed the relevant
241 part of the survey on Mechanical Turk. Three participants had to be excluded because they
242 indicated that we should not use their data, resulting in a final sample of $N = 312$ (55%
243 women, 44% men, 1% other or not disclosed; Mean age = 37.6 years, age range: 18 – 76).

244 *Type of technology and frequency of use.* The most frequently mentioned technology was
245 Facebook (42.7%), followed by other social media (Twitter 7.4%, Instagram 7.1%, YouTube
246 3.5%), Fitbit (3.2%), Microsoft Word or Excel (2.6%, 1.9%) and various other technologies
247 among others Mechanical Turk, web browser, Amazon Alexa, digital games, and mobile
248 apps. A vast majority of participants indicated that they used the technology multiple
249 times a day (84.6%). Almost 44% indicated that they used the technology six or more times
250 a day.

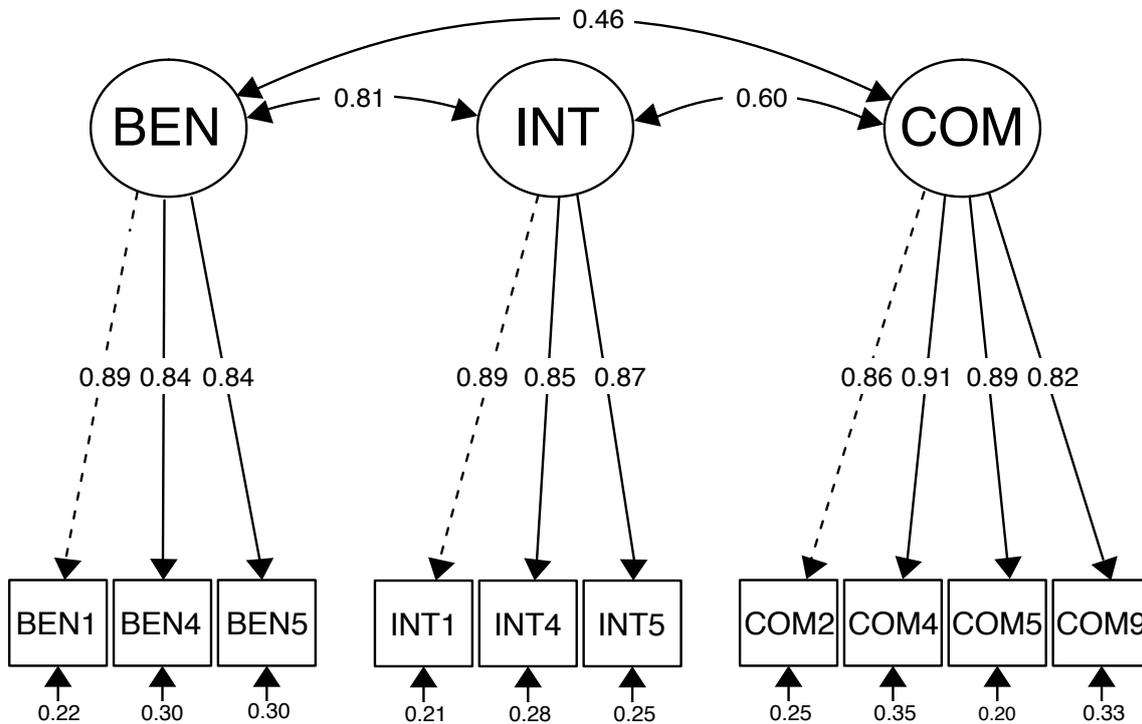


Figure 2: Measurement model of the TrustDiff in Study 2 with standardized loadings. Dotted lines indicate loadings that were constrained to one [$\chi^2(32) = 32.500, p = .442, \chi^2/df = 1.02, CFI = 1.000, SRMR = .027, RMSEA = .007, PCLOSE = .996$]

251 *5.2. Results and Discussion*

252 As a test of the three-dimensional factor structure, a confirmatory factor analysis was
 253 conducted using the lavaan package (0.5-23.1097) for R. All items were specified to load on
 254 their designated factor, and the loading of the first item was constrained to one. Multivari-
 255 ate normality was not given (Mardia tests: $\chi_s^2 = 2474.4, p < .001; Z_k = 50.6, p < .001$),
 256 therefore we used a robust maximum likelihood estimation method with Huber-White stan-
 257 dard errors and a Yuan-Bentler based scaled test statistic. Results of the CFA including all
 258 14 items suggested that the proposed model does adequately but not perfectly fit the data
 259 [$\chi^2(74) = 140.530, p < .001, \chi^2/df = 1.89, CFI = .971, SRMR = .047, RMSEA = .054,$
 260 $PCLOSE = .279$]. All loadings of the latent factors on their designated items exceeded
 261 .80 except for item BEN2 (.67). Investigation of modification indices suggested covariance

262 between items COM3 and COM4 as well as COM3 and COM5 and a loading of Compe-
263 tence on INT4 to improve model fit. However, since the goal was to create an economic
264 scale for user trust with three subscales, certain items have been removed instead of allow-
265 ing cross-loadings for a better model fit. Thus, item BEN2 (malicious - benevolent) was
266 removed because of low loadings of the Benevolence factor, INT4 (unbelievable - believable)
267 was removed to reduce a possible influence of Competence on Integrity, and Item COM5
268 (uninformed - informed) was primarily removed on theoretical grounds. The aspect of how
269 informed a vendor of a product is seems to be less related to other aspects of competence
270 such as capability, qualifications, and resources. The item COM3 (unskilled - skillful) was
271 removed because it has too much statistical and theoretical overlap with item COM4 (un-
272 qualified - proficient).

273 The final scale was reduced to 10 items, measuring the three related but distinct dimen-
274 sions and showed excellent psychometric properties [$\chi^2(32) = 32.500, p = .442, \chi^2/df = 1.02,$
275 $CFI = 1.000, SRMR = .027, RMSEA = .007, PCLOSE = .996$]. Descriptive statistics
276 of the final 10-item TrustDiff are depicted in Table 6 and the measurement model is shown
277 in Figure 2. Internal consistency of the three subscales was high ($\alpha_{Ben} = .85, \alpha_{Int} = .90,$
278 $\alpha_{Com} = .91$).

279 Results of these two confirmatory factor analysis showed that the questionnaire could be
280 improved and shortened without losing reliability. The final model for the 10-item TrustDiff
281 presents an excellent fit with high internal consistency. In a next step, an experiment was
282 conducted to investigate criterion validity of the scale.

283 6. Study 3

284 The goal of Study 3 was to test whether the TrustDiff is able to differentiate between
285 two websites that were manipulated regarding their trust-related features.

286 6.1. Method

287 *Procedure and Materials.* As part of a larger research project, but unrelated to Study 1 or
288 Study 2, participants were asked to rate a mock online shop based on a screenshot provided.

Table 6: Descriptive statistics and Pearson correlations of all items for the final TrustDiff in Study 2.

	M	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. BEN1: ignoring – caring	4.43	1.39	–									
2. BEN4: insensitive – sensitive	4.27	1.38	0.74	–								
3. BEN5: inconsiderate – empathic	4.46	1.37	0.74	0.71	–							
4. INT1: dishonest – honest	4.78	1.47	0.64	0.59	0.61	–						
5. INT4: unbelievable – believable	5.06	1.50	0.59	0.56	0.56	0.76	–					
6. INT5: untruthful – truthful	4.68	1.45	0.65	0.59	0.59	0.78	0.72	–				
7. COM2: incompetent – competent	5.79	1.24	0.34	0.29	0.30	0.40	0.50	0.46	–			
8. COM4: unqualified – proficient	5.64	1.31	0.34	0.29	0.32	0.37	0.48	0.39	0.69	–		
9. COM5: incapable – capable	5.81	1.29	0.40	0.34	0.41	0.47	0.52	0.49	0.77	0.73	–	
10.COM9: inept – resourceful	5.79	1.34	0.36	0.27	0.30	0.36	0.48	0.42	0.72	0.68	0.72	–

Note. $N = 312$. All correlations are significant $p < .001$

289 The participants were randomly assigned into two groups. The first group were presented
 290 a screenshot of an online shop that included several trust-supporting elements (high trust)
 291 and the second group received a screenshot of an online shop that was lacking any trust-
 292 supporting elements (neutral) (see Figure 3). Graphic design, structure design, content
 293 design, and social-cue design elements were manipulated (see Table 7) according to the
 294 elements identified by Seckler et al. (2015) and Wang and Emurian (2005).

295 After examining the website screenshot for at least four seconds, participants were asked
 296 to fill in the TrustDiff, the Likert-type scale for trust by Flavián et al. (2006), and to rate
 297 visual appeal and perceived usability of the website. All measures were presented as in
 298 Study 1. Data collected for this part was used to assess if the TrustDiff can differentiate
 299 between high and neutral trustworthiness.

300 *Participants.* A total of 394 participants from the United states completed the relevant part
 301 of the survey on the crowdsourcing platform CrowdFlower. Data was cleaned with two
 302 attention check items which reduced the sample size to 258. Six additional participants had
 303 to be excluded because they indicated that we should not use their data, resulting in a final
 304 sample of $N = 252$ (71% women, 28% men, 1% other or not disclosed; Mean age = 39 years,

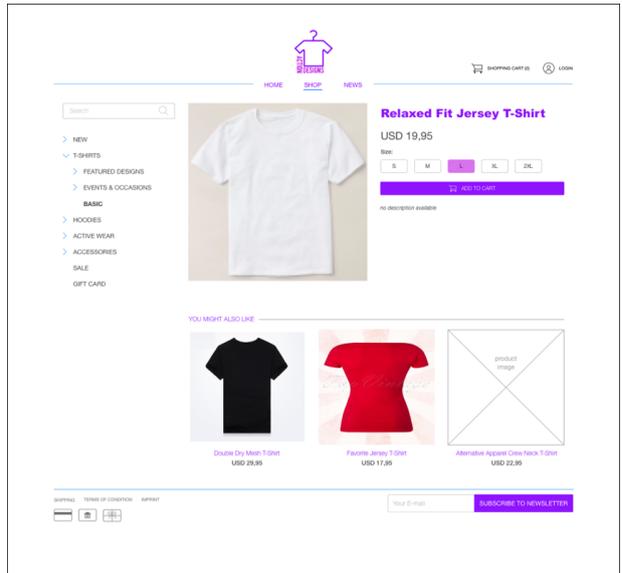
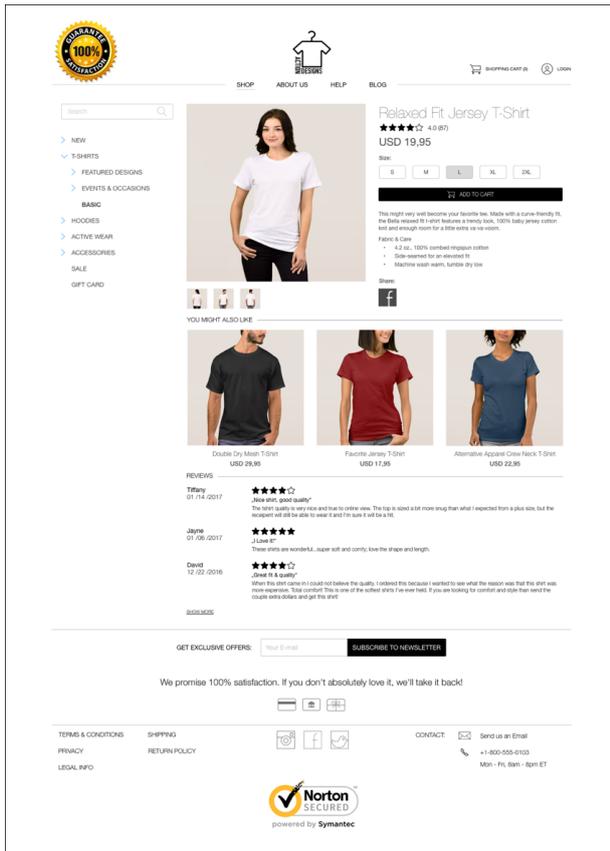


Figure 3: Mock online shop with trust-related features (left) and without (right) used in the experiment in Study 2. *Note to reviewers: a high-resolution image is included as a separate file at the end of this manuscript*

Table 7: Overview of the dimensions and respective features manipulated in the mock online shop.

Dimension	High-Trust	Neutral
Graphic design	Muted colours, high contrast	Bright colours, lower contrast
	Well-chosen and well-shot photographs	Inconsistent and missing photographs
Content design	Satisfaction guarantee	No satisfaction guarantee
	Links to more information in the footer, clearly readable	Hard to read or lacking information
	Link to the privacy policy	No link visible
	Seals of approval or third-party certificate	No seals of approval or third-party certificates
	Comprehensive, correct, and current product information	No product information
Social-cue design	Contact information for customer service in the footer	No contact information
	Users' reviews visible	Lack of users' reviews

305 age range: 18 – 78).

306 *Measures.* The 10 word-pairs of the final TrustDiff were included together with fifteen items
307 of the Trust scale developed by Flavián et al. (2006) and eighteen items of the VisAWI
308 measure for visual aesthetics Moshagen and Thielsch (2010)(Cronbach's $\alpha = .95$). Unlike
309 Study 1, only the overall score of the Trust scale by Flavián et al. (2006) was included in the
310 analysis (Cronbach's $\alpha = .96$). All three subscales of the TrustDiff showed excellent internal
311 consistency ($\alpha_{Ben} = .86$, $\alpha_{Int} = .90$, $\alpha_{Com} = .94$). As with Study 1 and Study 2, seven-point
312 scales were employed and the items were presented randomly.

313 6.2. Results

314 On average, participants viewed the websites for 1.47 minutes ($SD = 1.4$, min = 13.8
315 seconds, max = 14.08 minutes). No significant differences in viewing time (log-transformed)
316 were observed between the conditions, $t(246.88) = 0.073065$, $p = 0.9418$. All measured
317 deviated significantly from normal distribution, therefore Welchs's two samples t-test and
318 robust Wilcoxon rank-sum tests were conducted. Both tests led to the same conclusions for

Table 8: Descriptive statistics and results of Welch’s two samples t-test as an assessment of criterion validity of the TrustDiff.

		High trust (<i>n</i> = 128)		Neutral (<i>n</i> = 124)		t	df	<i>p</i>	<i>d</i>
		M	SD	M	SD				
TrustDiff	Benevolence	5.01	0.962	4.28	1.075	5.681	245.0	< .001	0.72
	Integrity	5.48	0.993	4.75	1.199	5.210	238.7	< .001	0.66
	Competence	5.58	1.011	4.47	1.455	6.989	218.6	< .001	0.89
	Total	5.37	0.899	4.50	1.176	6.577	230.1	< .001	0.84
Trust		5.11	0.947	4.20	1.254	6.470	228.8	< .001	0.82
VisAWI		4.91	1.094	3.91	1.167	7.037	247.7	< .001	0.89

Note. Total *N* = 252.

all measures, so we decided to list only the results of the Welch’s t-test. Criterion validity was investigated by comparing high trust condition with the neutral condition. As presented in Table 8, Welch’s two samples t-tests yielded significant differences between the conditions for all subscales of the TrustDiff and the total score ($t(230.1) = 6.577, p < .001, d = 0.84$). The Likert-type scale for trust Flavián et al. (2006) also showed a significant difference between the two conditions ($t(228.8) = 6.470, p < .001, d = 0.82$). The difference between both websites was even more pronounced for aesthetics which was generally rated lower by the participants ($t(247.7) = 7.037, p < .001, d = 0.89$).

7. General Discussion

The aim of this project was to develop and validate a scale measuring trust in online contexts using a semantic differential. Scale construction is an important step in confirmatory research because the quality of a measurement scale determines the extent to which empirical results are meaningful and accurate (Bhattacharjee, 2002).

The main contribution of the TrustDiff is two-fold: First, the semantic differential warrants a broad applicability for measuring user trust on the web. As discussed earlier, the

334 majority of existing trust questionnaires make use of the Likert-type items, which are mostly
335 tailored to the specific website measured, which makes it difficult to use these question-
336 naires in other research contexts (e.g., Lu et al., 2012; McKnight et al., 2002). The pairs of
337 antonyms used in the TrustDiff however, comprise of adjectives which generally fit to any
338 context related to user trust on the web. Second, each item of the TrustDiff contains merely
339 two words (one item-pair), namely two contrary adjectives, which are easier to translate into
340 other languages than full sentences. The declarative statements used in Likert-scale items
341 from other trust scales (e.g., Bhattacharjee, 2002; Cho, 2006; Flavián et al., 2006; Gefen,
342 2002; McKnight et al., 2002) however are often complex and time-consuming to translate.
343 Taken together, the advantage of the TrustDiff over other trust scales is its broader and eas-
344 ier applicability in different contexts and languages, keeping the possible loss of reliability
345 and validity on a minimum level and its ability to measure different manifestations of trust
346 from a negative to a positive pole in one scale. International firms whose online-services are
347 available across numerous countries and different languages might profit from an universally
348 applicable trust scale. A company may lose a lot if they fail to assess consumer trust in
349 their services, especially when revenue structure depends on frequent and continuous user
350 transactions. Early identification of users with low trust levels may help to ensure their
351 retention by targeting them specifically with specialized interventions.

352 Based on existing literature 28 positive adjectives with up to 3 antonyms for the three
353 dimensions of trust (Benevolence, Integrity, and Competence) were generated. These items
354 were tested for appropriate linguistic and psychological bipolarity by an expert panel and
355 reduced to 20 item pairs. Results from factor analysis in Study 1 (N = 601) suggested a
356 14-item scale measuring three distinct but related dimensions of trust. The trust dimensions
357 of the 14-item TrustDiff were relatively highly correlated with a Likert-type trust scale and
358 less pronounced but still substantially correlated with perceived usability and aesthetics. In
359 Study 2 the 14-items questionnaire measurement model was tested with 312 participants
360 rating various frequently used technologies. Results of a confirmatory factor analysis sug-
361 gested several avenues for improvement which resulted in a 10-item scale for trust with good
362 psychometric properties. Moreover, the results of Study 3 show that the TrustDiff is sen-

363 sitive to websites with differences in trust-related features. The rating differences between
364 the two websites were between $d = 0.66$ and 0.89 , commonly interpreted as between mod-
365 erate to large (Cohen, 1977). Compared to existing questionnaires that are content specific
366 (e.g., McKnight et al., 2002) or developed in other languages (e.g., Flavián et al., 2006),
367 the TrustDiff can be applied in various context and has been tested with English-speaking
368 participants. From a practitioner standpoint, the 10-item TrustDiff can be applied with-
369 out modifications to assess customers' level of trust in an enterprise or service and may be
370 translated easily to other languages.

371 The three studies presented here entail an initial thorough validation of the TrustDiff.
372 Although the scale offers promising psychometric properties, the TrustDiff needs to be fur-
373 ther tested with various products and services in different contexts. However, the 10-item
374 scale showed very good psychometric properties with a large variety of technologies in Study
375 2. The structure of the TrustDiff found in Study 2 needs to be replicated in different cultural
376 contexts and with other languages than English. For this task, a semantic differential is ideal,
377 as less translation effort is needed compared to traditional Likert-type scales. However, it is
378 still essential to establish psychometric bipolarity and structural validity in other languages.
379 The TrustDiff could be used to investigate how different web design elements relate to the
380 different dimensions of trust or distrust, since the present questionnaire represents the con-
381 struct trust from a negative to a positive pole on three subscales. Furthermore, to build
382 a comprehensive picture of user's trust and trust-related behaviors, the TrustDiff could be
383 combined with measures of trust in a technology. Trust in a technology has been found to be
384 related to the intention to explore and use more features of this particular technology (Mck-
385 night et al., 2011). This vendor-technology trust distinction could be particularly helpful
386 to better understand their relative influence in the adoption of a technology, post-adoption
387 use and the abandonment of a technology. Ultimately, researchers could investigate the pre-
388 dictive power of the TrustDiff regarding the trust-related behavior of users and how it may
389 relate to antecedents of trust. For instance, interface language quality which is a major issue
390 in multilingual software projects (e.g., Bargas-Avila and Brühlmann, 2016) could influence
391 user's trust in vendors. Additionally, the wording of the TrustDiff is not exclusive to the

392 web context since many of the items might have face-validity in other settings. For instance,
393 the validity of the TrustDiff could be investigated in areas of interpersonal trust or off-line
394 buyer-seller relationships. No less promising would be an attempt to discover profiles based
395 on users' responses on the scale. This may allow researchers and practitioners to design and
396 evaluate trust-related interventions targeted at specific subgroups.

397 **8. Conclusion**

398 We present the development and validation of a semantic differential that helps to eval-
399 uate users' trust and potentially serve as a tool to investigate how user trust emerges. The
400 development and validation followed best practices and the scale is readily applicable to
401 a variety of research questions. The TrustDiff was tested with over 1000 participants and
402 showed good psychometric properties and high reliability. The semantic differential is easy-
403 to-use and easy-to-translate and thus a viable alternative to existing Likert-scale format
404 questionnaires for user trust.

405 **9. Acknowledgements**

406 Special thanks to Elisa Mekler. This work has been approved by the Institutional Review
407 Board of the Faculty of Psychology, University of Basel under the numbers D-003-17 and
408 M-003-17.

409 **10. References**

- 410 L. V. Casaló, C. Flavián, M. Guinalú, The role of security, privacy, usability and reputation in the devel-
411 opment of online banking, *Online Information Review* 31 (2007) 583–603.
- 412 J. W. Driscoll, Trust and participation in organizational decision making as predictors of satisfaction,
413 *Academy of Management Journal* 21 (1978) 44–56.
- 414 C. Moorman, R. Deshpande, G. Zaltman, Factors affecting trust in market research relationships, *Journal*
415 *of Marketing* (1993) 81–101.
- 416 J. B. Rotter, A new scale for the measurement of interpersonal trust, *Journal of Personality* 35 (1967)
417 651–665.

- 418 R. J. Lewicki, C. Brinseld, Measuring trust beliefs and behaviours, in: F. Lyon, G. Mollering, M. N. K.
419 Saunders (Eds.), *Handbook of research methods on trust*, Edward Elgar Publishing, Cheltenham, UK,
420 2012, pp. 29–39.
- 421 L. van der Werff, C. Real, T. Lynn, Individual trust and the internet, in: R. H. Searle, A.-M. I. Nienaber,
422 S. B. Sitkin (Eds.), *The Routledge Companion to Trust*, Routledge, Oxford, UK, 2018.
- 423 Y. D. Wang, H. H. Emurian, An overview of online trust: Concepts, elements, and implications, *Computers*
424 *in Human Behavior* 21 (2005) 105–125.
- 425 A. Bhattacharjee, Individual trust in online firms: Scale development and initial test, *Journal of Management*
426 *Information Systems* 19 (2002) 211–241.
- 427 J. Cho, The mechanism of trust and distrust formation and their relational outcomes, *Journal of Retailing*
428 82 (2006) 25–35.
- 429 C. Flavián, M. Guinalú, R. Gurrea, The role played by perceived usability, satisfaction and consumer trust
430 on website loyalty, *Information & Management* 43 (2006) 1–14.
- 431 D. Gefen, Reflections on the dimensions of trust and trustworthiness among online consumers, *ACM Sigmis*
432 *Database* 33 (2002) 38–53.
- 433 D. H. McKnight, V. Choudhury, C. Kacmar, The impact of initial consumer trust on intentions to transact
434 with a web site: A trust building model, *The Journal of Strategic Information Systems* 11 (2002) 297–323.
- 435 Y. Kim, R. A. Peterson, A meta-analysis of online trust relationships in e-commerce, *Journal of Interactive*
436 *Marketing* 38 (2017) 44–54.
- 437 T. Verhagen, B. van Den Hooff, S. Meents, Toward a better use of the semantic differential in is research:
438 An integrative framework of suggested action., *Journal of the Association for Information Systems* 16
439 (2015) 108–143.
- 440 S. C. Chen, G. S. Dhillon, Interpreting dimensions of consumer trust in e-commerce, *Information Technology*
441 *and Management* 4 (2003) 303–318.
- 442 R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, *Academy of*
443 *Management Review* 20 (1995) 709–734.
- 444 Y. Bart, V. Shankar, F. Sultan, G. L. Urban, Are the drivers and role of online trust the same for all web
445 sites and consumers? A large-scale exploratory empirical study, *Journal of Marketing* 69 (2005) 133–152.
- 446 B. J. Corbitt, T. Thanasankit, H. Yi, Trust and e-commerce: A study of consumer perceptions, *Electronic*
447 *Commerce Research and Applications* 2 (2003) 203–215.
- 448 M. K. Lee, E. Turban, A trust model for consumer internet shopping, *International Journal of Electronic*
449 *Commerce* 6 (2001) 75–91.
- 450 S. L. Jarvenpaa, N. Tractinsky, L. Saarinen, Consumer trust in an internet store: A cross-cultural validation,
451 *Journal of Computer-Mediated Communication* 5 (1999) 0–0.

452 D. H. McKnight, V. Choudhury, C. Kacmar, Developing and validating trust measures for e-commerce: An
453 integrative typology, *Information Systems Research* 13 (2002) 334–359.

454 P. A. Pavlou, D. Gefen, Building effective online marketplaces with institution-based trust, *Information*
455 *Systems Research* 15 (2004) 37–59.

456 J. Lu, L. Wang, L. A. Hayes, How do technology readiness, platform functionality and trust influence C2C
457 user satisfaction?, *Journal of Electronic Commerce Research* 13 (2012) 50–69.

458 W. W. Chin, N. Johnson, A. Schwarz, A fast form approach to measuring technology acceptance and other
459 constructs, *MIS Quarterly* (2008) 687–703.

460 O. Friborg, M. Martinussen, J. H. Rosenvinge, Likert-based vs. semantic differential-based scorings of posi-
461 tive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience,
462 *Personality and Individual Differences* 40 (2006) 873–884.

463 D. I. Hawkins, G. Albaum, R. Best, Stapel scale or semantic differential in marketing research?, *Journal of*
464 *Marketing Research* 11 (1974) 318–322.

465 J. Wirtz, M. C. Lee, An examination of the quality and context-specific applicability of commonly used
466 customer satisfaction measures, *Journal of Service Research* 5 (2003) 345–355.

467 S. Van Auken, T. E. Barry, An assessment of the trait validity of cognitive age measures, *Journal of*
468 *Consumer Psychology* 4 (1995) 107–132.

469 D. Gefen, E. Karahanna, D. W. Straub, Trust and TAM in online shopping: An integrated model, *MIS*
470 *Quarterly* 27 (2003) 51–90.

471 I. B. Hong, H. Cho, The impact of consumer trust on attitudinal loyalty and purchase intentions in b2c
472 e-marketplaces: Intermediary trust vs. seller trust, *International Journal of Information Management* 31
473 (2011) 469–479.

474 J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated
475 systems, *International Journal of Cognitive Ergonomics* 4 (2000) 53–71.

476 M. Koufaris, W. Hampton-Sosa, The development of initial trust in an online company by new customers,
477 *Information & Management* 41 (2004) 377–397.

478 J. C. McCroskey, J. J. Teven, Goodwill: A reexamination of the construct and its measurement, *Communi-*
479 *ications Monographs* 66 (1999) 90–103.

480 D. C. Rieser, O. Bernhard, Measuring trust: The simpler the better?, in: *Proceedings of the 2016 CHI*
481 *Conference Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA,
482 2016, pp. 2940–2946.

483 M. C. Howard, R. C. Melloy, Evaluating item-sort task methods: The presentation of a new statistical
484 significance formula and methodological best practices, *Journal of Business and Psychology* 31 (2016)
485 173–186.

486 M. Hassenzahl, M. Burmester, F. Koller, Attrakdiff: Ein fragebogen zur messung wahrgenommener hedo-
487 nischer und pragmatischer qualität, in: Mensch & Computer 2003, Springer, 2003, pp. 187–196.

488 M. Moshagen, M. T. Thielsch, Facets of visual aesthetics, International Journal of Human-Computer Studies
489 68 (2010) 689–709.

490 K. Finstad, The usability metric for user experience, Interacting with Computers 22 (2010) 323 – 327.

491 G. D. Hutcheson, N. Sofroniou, The multivariate social scientist: Introductory statistics using generalized
492 linear models, Sage, 1999.

493 A. Field, Discovering statistics using IBM SPSS statistics, Sage, 2013.

494 M. C. Howard, A review of exploratory factor analysis decisions and overview of current practices: What
495 we are doing and how can we improve?, International Journal of Human-Computer Interaction 32 (2016)
496 51–62.

497 M. Seckler, S. Heinz, S. Forde, A. N. Tuch, K. Opwis, Trust and distrust on the web: User experiences and
498 website characteristics, Computers in Human Behavior 45 (2015) 39–50.

499 J. Cohen, Statistical Power Analysis for the Behavioral Sciences (Revised Edition), Academic Press, 1977.

500 D. H. Mcknight, M. Carter, J. B. Thatcher, P. F. Clay, Trust in a specific technology: An investigation of
501 its components and measures, ACM Trans. Manage. Inf. Syst. 2 (2011) 12:1–12:25.

502 J. A. Bargas-Avila, F. Brühlmann, Measuring user rated language quality: Development and validation of
503 the user interface language quality survey (LQS), International Journal of Human-Computer Studies 86
504 (2016) 1–10.



Search

- > NEW
- > T-SHIRTS
- > FEATURED DESIGNS
- > EVENTS & OCCASIONS
- BASIC
- > HOODIES
- > ACTIVE WEAR
- > ACCESSORIES
- SALE
- GIFT CARD



Relaxed Fit Jersey T-Shirt
 ★★★★★ 4.0 (87)
 USD 19.95

Size:

ADD TO CART

This might very well become your favorite tee. Made with a curve-friendly fit, the Bella relaxed fit t-shirt features a trendy look, 100% baby jersey cotton knit and enough room for a little extra wa-ya-voom.

- Fabric & Care
- 4.2 oz., 100% combed ring-spun cotton
 - Side-seamed for an elevated fit
 - Machine wash warm, tumble dry low

Share:



YOU MIGHT ALSO LIKE



Double Dry Mesh T-Shirt
 USD 29.95



Favorite Jersey T-Shirt
 USD 17.95



Alternative Apparel Crew Neck T-Shirt
 USD 22.95

REVIEWS

Tiffany
 01/14/2017

★★★★★
 "Nice shirt, good quality"

The shirt quality is very nice and true to online view. The top is sized a bit more snug than what I expected from a plus size, but the recipient will still be able to wear it and I'm sure it will be a fit.

Jayne
 01/06/2017

★★★★★
 "I Love it!"

These shirts are wonderful...super soft and comfy, love the shape and length.

David
 12/22/2016

★★★★★
 "Great fit & quality"

When this shirt came in I could not believe the quality. I ordered this because I wanted to see what the reason was that this shirt was so popular. The quality of the fabric and the colors are great. I've ordered 3 more. If you are looking for comfort and style then send the couple extra dollars and get the shirt.

SECURE

GET EXCLUSIVE OFFERS:

Your E-mail

SUBSCRIBE TO NEWSLETTER

We promise 100% satisfaction. If you don't absolutely love it, we'll take it back!



TERMS & CONDITIONS
 SHIPPING
 RETURN POLICY
 PRIVACY
 LEGAL INFO

CONTACT: Send us an Email
 +1-800-555-0103
 Mon - Fri, 8am - 5pm ET



powered by Symantec



Search

- > NEW
- > T-SHIRTS
- > FEATURED DESIGNS
- > EVENTS & OCCASIONS
- BASIC
- > HOODIES
- > ACTIVE WEAR
- > ACCESSORIES
- SALE
- GIFT CARD



Relaxed Fit Jersey T-Shirt
 USD 19.95

Size:

ADD TO CART

no description available

YOU MIGHT ALSO LIKE



Double Dry Mesh T-Shirt
 USD 29.95



Favorite Jersey T-Shirt
 USD 17.95



Alternative Apparel Crew Neck T-Shirt
 USD 22.95

SHIPPING TERMS OF CONDITION



SUBSCRIBE TO NEWSLETTER

Your E-mail

Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data
Quality in Crowdsourced Samples

Florian Brühlmann¹, Serge Petralito, Lena F. Aeschbach, and Klaus Opwis

Center for Cognitive Psychology and Methodology

University of Basel

Missionsstrasse 62a

CH-4055 Basel, Switzerland

Affiliation

¹ Corresponding author. Tel.: + 41 (0)61 207 06 66

Abstract

Research in various academic fields relies increasingly on online samples. With the advent of crowdsourcing platforms, online data collection has become more popular than ever, although concerns have been raised recently. These concerns regard the data quality of these samples and the possible adverse effects of poor data on experimental manipulations and scale properties. Presently, research on carelessness in crowdsourced surveys is scarce. Therefore, the goal of this study (N = 394) was to systematically identify careless and inattentive behavior in a crowdsourced sample by applying various measures and methods for detecting carelessness. Results revealed that approximately half of all participants were inattentive in the online survey. Furthermore, carelessness and inattentive behavior appear highly task-dependent, because correlations between open answer quality and other measures were rather low. Thus, based on a predictive model and ease of interpretation, we recommend assessing the data quality of crowdsourced samples with a self-reported single item, one or multiple attention checks (such as an Instructed Response Item (IRI)), a LongString analysis, and a task-specific measure. This combination of detection methods accurately predicted careless participants, and excluding these participants increased the effect size in an experiment included in the survey.

Keywords: Inattentive responding; Careless responding; Crowdsourcing; Response patterns; Open answer; Latent profile analysis

Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples

Introduction

Online surveys have become a standard method of data collection in various fields, such as in recent psychological research (Gosling & Mason, 2015) and market research (Comley, 2015). Whereas in 2003 and 2004 only 1.6% of articles published in APA journals used the Internet (Skitka & Sargis, 2006), Gosling and Mason (2015) stated just a few years later that “studies that use the Internet in one way or another have become so pervasive that reviewing them all would be impossible” (p. 879). Moreover, this method covers virtually all areas of psychology. Online data collection has numerous advantages over laboratory studies: lower infrastructure costs (no laboratory infrastructure or individual time slots are needed), faster and cheaper data collection (Casler, Bickel, & Hackett, 2013; de Winter, Kyriakidis, Dodou, & Happee, 2015), more extensive distribution of the study, and lower hurdles for participation (Kan & Drumme, 2018). One of the most popular recruitment methods for participants in online studies for psychological research is the use of crowdsourcing services, such as Amazon’s Mechanical Turk (MTurk) or FigureEight (formerly known as CrowdFlower). Regarding MTurk, approximately 15’000 published articles used this crowdsourcing platform between 2006 and 2014 for their data collection (J. Chandler & Shapiro, 2016; Kan & Drumme, 2018). On these platforms, various small tasks are offered in exchange for money to “crowd workers”. All the advantages of other online data collection methods, such as cost- and time-effectiveness, also apply to crowdsourcing platforms (Kan & Drumme, 2018). Additionally, crowdsourcing platforms offer a more diverse population compared to typically homogenous samples from psychological studies (Kan & Drumme, 2018): In the case of MTurk, these workers are composed of a demographic containing more than 500’000 individuals from 190 countries (Paolacci & Chandler, 2014). While concerns considering the generalizability and validity of crowdsourced online samples have been discussed (Kan & Drumme, 2018), Gosling and Mason (2015) also reported that the mean and range of ages from an MTurk-sample are more representative of the general US population than a sample merely consisting of undergraduate students. Moreover, in comparison to online samples

recruited on social media platforms, some crowdsourced samples were found to have a higher diversity in terms of age, cultural, and socioeconomic factors (Casler et al., 2013), and more balanced gender ratios (de Winter et al., 2015). Furthermore, Kan and Drummey (2018) stated that MTurk is a viable alternative to traditional methods of data collection, because many studies showed similar patterns of findings in their crowdsourced data when compared to results using traditional approaches of data collection (Kan & Drummey, 2018, p. 244).

However, given the increased distance between researchers and participants in online studies, and the possible influence of distractions in an uncontrolled setting, data collected online may suffer from bad quality stemming from inattentiveness and other forms of deceptive behavior.

Participant carelessness or inattentiveness (Meade & Craig, 2012), have recently received increased attention from various researchers regarding their reasons, effects, detection, and prevention (Maniaci & Rogge, 2014; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). Although carelessness may also occur in laboratory studies, the problem seems especially common within online samples because survey administrations are often unproctored (Cheung, Burns, Sinclair, & Sliter, 2017; Fleischer, Mead, & Huang, 2015). Maniaci and Rogge (2014) went further, claiming that the current “replication crisis” in psychology may be related to careless respondents who take part in online surveys with insufficient attention. Regarding crowdsourced samples, concerns about the data quality have also been raised, as these workers are usually non-naive participants with possibly deceptive behavior. This tendency is exacerbated by the incentive-structure of these platforms. Further, the responses are often conducted in uncontrolled and possibly distracting environments (J. Chandler, Mueller, & Paolacci, 2014; J. Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Kan & Drummey, 2018; Peer, Brandimarte, Samat, & Acquisti, 2017; Stewart, Chandler, & Paolacci, 2017).

Causes and effects of carelessness

In the present study, we primarily focus on participant inattention or careless response. Other forms of invalid responding and deceptive behavior (such as social desirability and faking responses), also decrease data quality, but may have different causes and effects (Maniaci &

Rogge, 2014; McKay, Garcia, Clapper, & Shultz, 2018). Participant inattention might have many sources, one of them being the anonymity of computer-based surveys, which can result in a lack of accountability (Douglas & McGarty, 2001; Lee, 2006; Meade & Craig, 2012). Further important factors affecting carelessness in survey data are respondent interest, length of survey, social contact, and environmental distraction (Meade & Craig, 2012). Extrinsic motivation might also account for carelessness, such as when participants are paid for their answers. Gadiraju, Kawase, Dietze, and Demartini (2015) found that some participants recruited via crowdsourcing services employ strategies to minimize their invested time or effort in return for participation compensation. In these cases, careless responding and subsequent poor data quality emerge from crowdworkers who are solely interested in receiving their payment as fast as possible without providing valid data for the researcher. Furthermore, Niessen et al. (2016) observed that students also strove to complete surveys as quickly as possible in exchange for course credits. Aside from these external factors, a study conducted by McKay et al. (2018) found that careless responding is strongly related to malevolent personality traits, whereas its connection to benevolent traits was less pronounced.

Base rate estimates for bad online data quality stem from different concepts of invalid responding and different sources for online data collection: Recent research has estimated that, depending on the method, between 10% to 12% of participants in an online survey exhibit an answering behavior described as insufficient effort responding or careless responding (Meade & Craig, 2012). In a more heterogeneous sample, Maniaci and Rogge (2014) found that between 3% to 9% of participants respond carelessly. In an online survey with students from a university in the United States, Ward, Meade, Allred, Pappalardo, and Stoughton (2017) showed that 23% of the participants were flagged by at least one IRI. Collecting online data from a Facebook sample, Dogan (2018) estimated careless responding in 40.7% to 59.8% of the sample, depending on the measure used to detect careless behavior. For a crowdsourced sample on MTurk, Kan and Drummey (2018) found that between 21.8% and 55.8% of the sample (depending on eligibility requirements), proved deceptive and provided false data. However, it is important to note that Kan and Drummey (2018) did not assess carelessness or inattentive behavior. They solely refer to deceptive behavior concerning screening

requirements on Amazon's MTurk, which emerge under certain eligibility constraints. Carelessness in Hauser and Schwarz (2016) was assessed merely by using an instructional manipulation check, resulting in highly volatile estimates from 4% to 74.5%, depending on the exact method used. Instructional manipulation checks have also been criticized for being too restrictive, as partially skipping instructions does not automatically mean that participants are inattentive (Maniaci & Rogge, 2014). In another study assessing carelessness in a crowdsourced sample, Peer et al. (2017) found that only 27% of all participants in a FigureEight-sample passed all attention checks, and approximately 18% failed in all of them. While these numbers provide some valuable insights for assessing careless behavior in crowdsourced samples, the study did not include other carelessness measures. Therefore, it only identified one behavioral form of inattention or carelessness. Consequently, these alarmingly high estimates for bad data quality stemming from carelessness or other deceptive forms of behavior vary greatly between studies, methods, and recruitment methods. However, even a seemingly small number of careless responses can have serious consequences, such as failed replications (Oppenheimer, Meyvis, & Davidenko, 2009) or false-positives (Huang, Liu, & Bowling, 2015). Furthermore, careless responding may cause failed manipulations when instructions are not carefully read (Maniaci & Rogge, 2014), lower internal consistency of validated scales (Maniaci & Rogge, 2014), and problems in questionnaire development and item analysis (Johnson, 2005). Additionally, it can lead to problems in investigating questionnaire dimensionality (Kam & Meyer, 2015). However, estimates for careless behavior in crowdsourced samples remain unknown. All the aforementioned research examined academic participant pools or mixed types of online data (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012), or the studies only applied one measure to determine carelessness in a crowdsourced sample (Dogan, 2018; Hauser & Schwarz, 2016; Peer et al., 2017). Therefore, it was concluded there is a lack of a systematic analysis of careless behavior on crowdsourced platforms using various carelessness detection methods.

Recently, most of the attention of empirical research has been given to the discovery of carelessness (see Curran, 2016, for a review). The screening methods can be divided into two groups. The first group is the planned implementation of special items or scales to screen

carelessness. For example, Bogus Items (Meade & Craig, 2012), IRIs (Curran, 2016), and instructional manipulation checks (Oppenheimer et al., 2009). The second group of detection methods can be described as post hoc measures. These include the examination of response time, multivariate outliers, and (in-) consistency indices. These do not require special items, but an elaborate analysis after data collection. There are a variety of different methods available, but we will focus on those recommended in the recent literature (Curran, 2016).

Aim of the present study

Thus, the aim of the present study was to analyze the data quality of a crowdsourced online sample, based on various recommended methods for assessing careless behavior. This would address the limited variety of methods used in existing research about carelessness on crowdsourcing platforms.

Another open question revolves around the task-dependence of carelessness, and whether different methods embedded in different tasks capture different participants, or whether they stay careless for most of the study. As stated by Kan and Drummey (2018), it remains unclear in what way the duration or engagement level of a task impacts deceptive or careless behavior. Besides Likert-type scales, open questions (for example) are an extensively used method for capturing qualitative data. However, it is unknown how the quality of the answers given to such questions relates to carelessness.

To address these problems, the present study aims toward a better understanding of careless and inattentive behavior on crowdsourcing platforms (and the task-dependence of this phenomenon), by assessing the data quality with various detection methods. Moreover, we aim to provide pragmatic recommendations for ensuring survey data quality in research with crowdsourced samples. Based on these aims, we derived the following research questions:

Research Question 1: How prevalent is careless responding in samples from crowdsourcing platforms, based on various detection methods for carelessness?

Research Question 2: How are task-specific measures of carelessness (such as open-ended questions) related to planned detection methods and post hoc methods?

Research Question 3: Based on our findings, which methods are most applicable for

identifying carelessness in a crowdsourced sample?

Method

Data collection

The present study was conducted using a crowdsourced sample from FigureEight (CrowdFlower). Especially outside the U.S., FigureEight is a viable choice for crowdsourcing, as Amazon's MTurk has (for a long time) required requesters to have a US-address.

FigureEight is accessible from Europe and other regions outside the USA, and provides access to millions of contributors (Van Pelt & Sorokin, 2012). The crowdsourcing platform is a well-established tool to gather participants for online-surveys, as shown by over 4600 hits on Google Scholar (21.03.2018, Keyword: CrowdFlower).

Data and analysis code used in this study is available at

https://osf.io/9vjur/?view_only=9ed1707502684f89be168d358f5cd695

(anonymized for peer review).

Procedure

After providing consent, participants were asked to recall a recent negative experience with an online store. In particular, participants were asked to respond to two questions 1) what exactly caused this experience to be negative and 2) how this affected their online shopping habits.

We instructed participants to respond in free text with as much detail as possible, with complete sentences, and with at least 50 words. Next, 10 items of the positive and negative affect schedule (PANAS; Watson, Clark, & Tellegen, 1988), 23 items of the AttrakDiff2 (Hassenzahl, Burmester, & Koller, 2003), and 24 items measuring psychological need satisfaction adapted from Sheldon, Elliot, Kim, and Kasser (2001) were presented. This type of critical incident method is a common procedure in user experience research (e.g., Tuch, Schaik, & Hornbæk, 2016). After this first block of questions, participants were randomly allocated to be shown either a high trust or low trust mockup of a website. The website was manipulated according to the trust supporting elements identified by Seckler, Heinz, Forde, Tuch, and Opwis (2015). This setting was chosen to conduct a plausible experiment in user

experience research that was thematically related to the rest of the study. After this, participants were asked to complete 16 items of a Likert-type scale for trust in websites (Flavián, Guinalú, & Gurrea, 2006). The goal of this section was to examine the effects of excluding data from careless participants on effect sizes and p-values in a group comparison. On the next page, participants rated the visual aesthetics of the website mock-up with 18 items (VisAWI, Moshagen and Thielsch (2010)). Following this section, the big five personality types were assessed with 44 items of the Big Five Inventory (BFI) (John & Srivastava, 1999). All post hoc detection methods of carelessness were investigated using this scale. On the last page of the survey, participants completed demographic information and a scale on self-reported careless responding (as in Maniaci & Rogge, 2014) and a self-reported single item (SRSI UseMe) (Meade & Craig, 2012). Finally, all participants were given a completion code.

Measures

All post hoc detection methods of carelessness were applied to the 44 items of the BFI in the last part of the questionnaire. We decided to focus on the BFI because it is multidimensional with a sufficient length to calculate various indices, and it is comparable with other studies in this field (Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012). The data of the other questionnaires used in this study were not subject to further analysis except for the trust scale by Flavián et al. (2006), which was used as a dependent variable in the experiment.

Planned detection methods

The wording of the self-reported responding tendencies scale, the Bogus Item, and the IRI incorporated in the study is presented in Table 1.

Self-reported responding tendencies. Following demographic questions, ten items based on Maniaci and Rogge (2014) were used to assess general tendencies in responding. Although excluding participants based on self-reported responding tendencies has been found to improve data quality significantly (Aust, Diedenhofen, Ullrich, & Musch, 2013), these items are also easily detected and prone to manipulation and dishonest answers. Three items were used to measure self-reported careless responding ($\alpha = .84$), two items to measure

self-reported patterned responding ($\alpha = .88$), three items to assess self-reported rushed responding ($\alpha = .83$), and two items assessing self-reported skipping of instructions ($\alpha = .68$). All items are presented in Table 1. Items were rated on a 7-point scale (1 = never, 4 = approximately half the time, 7 = all of the time), and responses were averaged ensuring high scores reflected more problematic responding.

Applying the cutoff used by Maniaci and Rogge (2014), answers of 4 or higher were flagged. Additionally, self-indicated data usage was assessed using the SRSI UseMe.

Attention checks. We employed two attention check items in the questionnaire following the Infrequency Approach (Huang, Curran, Keeney, Poposki, & DeShon, 2012), which entails including items to which all careful respondents should respond to in the same (or similar) fashion. One measure we applied was the Bogus Item similar to Meade and Craig (2012), which are items that are very unlikely for participants to agree with. The Bogus Item was located within the BFI (see Table 1). Participants who did not select "strongly disagree" or "slightly disagree" were thus flagged as inattentive. The other attention check item was an IRI similar to Meade and Craig (2012) and Curran (2016). According to Meade and Craig (2012) the IRI has several advantages over Bogus Items, as they are easier to create, have a singular correct answer, and therefore provide an obvious metric for scoring. Furthermore, they offer a clear interpretation and are not prone to humorous answers, which is a problem with the Bogus Item. The IRI (see Table 1) was placed within the items of the trust scale by Flavián et al. (2006). Participants who nevertheless answered this question were flagged.

Post hoc detection methods

Response Time. One simple post hoc measure to assess careless responding is to measure participant overall response time. The concept is that inattentive or careless respondents will be noticeable through unusually short or long completion times. Although this measure is easily applicable in any online survey, the issue of what constitutes an acceptable range of completion times must be decided individually for each question (Curran, 2016).

LongString Index. The LongString Index acts as an invariability measure, which assesses the number of same answers given in sequence. Careless participants who might select the

Table 1

The items of the self-reported responding tendencies scale (Maniaci & Rogge, 2014) and planned detection items included in the study. Self-report answer options ranged from 1 (never), over 4 (about half of the time) to 7 (all the time). The Bogus Item was included in the BFI where answers between 1 (disagree strongly) and 5 (agree strongly) were possible. The IRI was included in the trust scale that was used as the dependent variable of the experiment.

Measure	Item
Self-report	[How often do you...]
Careless responding	<ol style="list-style-type: none"> 1. Read each question 2. Pay attention to every question 3. Take as much time as you need to answer the questions honestly
Patterned responding	<ol style="list-style-type: none"> 4. Make patterns with the responses to a block of questions 5. Use the the same answer for a block of questions one the same topic [rather than reading each question]
Rushed responding	<ol style="list-style-type: none"> 6. Answer quickly without thinking 7. Answer impulsively without thinking 8. Rush through the survey
Skipping of instructions	<ol style="list-style-type: none"> 9. Skim the instructions quickly 10. Skip over parts of the instruction
SRSI UseMe	In your honest opinion, should we use your data in our analyses in this study? (Do not worry, this will not affect your payment, you will receive the payment code either way.)
Bogus Item	[I see myself as someone who ...] Did not read this statement
IRI	I read instructions carefully. To show that you are reading these instructions, please leave this question blank.

same answer for equal or greater than half the length of the total scale will be excluded from the sample (Curran, 2016; Huang et al., 2012). Curran (2016) recommended LongString analysis to identify some of the worst respondents that would otherwise be missed, although the measure can easily be deceived. The LongString Index in this study was calculated for the BFI following the procedure described in Meade and Craig (2012).

Odd-even consistency. To assess the Odd-even consistency (OEC), each individual's responses on each unidimensional subscale are split into responses to even and to uneven items (Curran, 2016). In the present case, this was implemented for each of the five dimensions of the BFI (Openness, Extraversion, Agreeableness, Conscientiousness, and Neuroticism). Reverse coded items must be recorded before calculating this measure. The responses to the even and uneven items are then averaged separately, ensuring each participant receives a score based on the even and the uneven items for each subscale of one larger scale. The individual correlation of these two vectors acts as a score of consistency. An important limitation is that this correlation is constrained by the number of subscales and the number of items in a scale. The OEC in this study was assessed for the BFI based on the procedure described by Meade and Craig (2012). Following Curran's (2016) recommendation, any correlation below 0 was flagged.

Resampled individual reliability. Curran (2016) proposed a more general conceptualization of the OEC measure – Resampled individual reliability (RIR). Here, the basic concept is that items that should measure the same construct should correlate positively within individuals. However, instead of limiting this idea to odd and even items, Curran (2016) suggests creating two halves of each subscale randomly without replacement. The individual correlation of these two vectors acts as a score of consistency. This process is then repeated several times (resampling). This is a new measure that was included in the present study and, to the best of our knowledge, has never been empirically examined. Following Curran's (2016) recommendation for the OEC, any correlation below 0 was flagged.

Person-total correlation. The measure of Person total correlation (PTC) describes the correlation of a participant's answers to each of the items of a scale, with the means of these items based on the whole sample (Curran, 2016). This measure relies on the assumption

that a large majority of the sample responded attentively, thus this measure may be problematic in situations where a large number of careless respondents are expected. Because this measure has currently not been empirically examined, no widely accepted cutoff value for this correlation exists. However, as recommended by Curran (2016), participants with a negative PTC were flagged.

Open answer quality. A priori criteria for the quality rating of open answers predominantly originates from the studies conducted by Holland and Christian (2009) and (Smyth, Dillman, Christian, & McBride, 2009). The following indicators for calculating an open answer quality index were taken into consideration: 1. Whether participants provided a thematically substantive response. 2. If a minimum of 50 words was provided (as instructed). 3. If participants provided answers in complete sentences (as instructed). 4. The number of subquestions answered (as instructed). 5. The number of subquestions further elaborated. A detailed description of how the open answer quality index was created is presented in the Appendix. The third author coded all experience reports. To ensure inter-rater reliability, the second author coded a random subset of 100 open-ended answers. Because two fixed raters rated a randomly selected subset, ICC3 was used (Koo & Li, 2016). Inter-rater agreement of each category was between moderate (Complete Sentences, $ICC3 = .80$), good (Substantive Response, $ICC3 = .78$; Number of Subquestions Elaborated, $ICC = .84$) and excellent (Number of Subquestions Answered, $ICC3 = .94$). Inter-rater agreement for the overall answer quality index was excellent $ICC3 = .96$, with a 95% confidence interval from .94 to .97 ($F(99,99) = 51, p < .001$).

Results

In this section, we first report on each group of carelessness detection methods separately, and then investigate how they relate to answer quality. Table 2 presents an overview of the number of participants flagged by each method.

Planned detection methods

Self-reported responding tendencies. Participants relatively frequently indicated that they engaged in problematic responding tendencies. Applying the cutoff used by Maniaci and

Table 2

Descriptive statistics for all detection methods used in the study. Self-report includes problematic responding tendencies as well as the SRSI UseMe item.

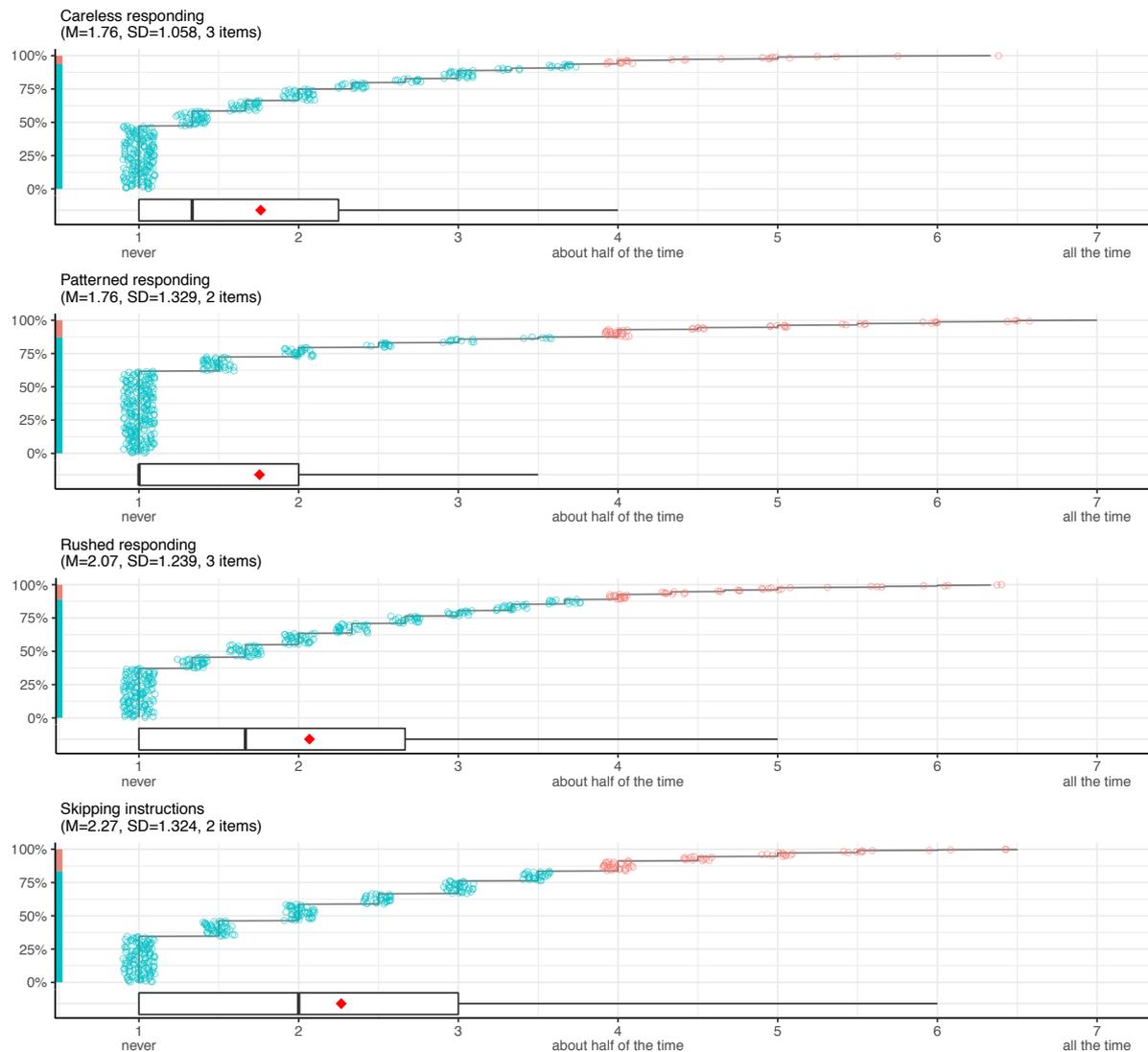
	Mean	SD	Min	Max	No. Flagged	%
Planned detection						
Self-report					106	26.90
Bogus Item					92	23.35
Instructed Response Item					96	24.37
Response time	16.71	9.22	3.93	61.15		
Post hoc detection						
LongString	6.63	9.15	0	44	25	6.35
Odd-even consistency	0.61	0.43	-1	1	63	15.99
Resampled individual reliability	0.56	0.39	-0.82	0.99	63	15.99
Person-total correlation	0.38	0.32	-0.47	0.88	74	18.78
Answer quality					100	25.38
Total (flagged by at least one method)					233	59.14

Note. Total $N = 394$

Rogge (2014), answers with 4 or higher were flagged. Thus, we flagged 25 careless respondents (6.6%), 50 pattern-respondents (12.7%), 44 rushed-respondents (11.2%), and 65 (16.5%) participants for skipping instructions. As depicted in Figure 1, skipping instructions was admitted most frequently ($M = 2.27$) followed by rushed responding ($M = 2.07$). However, there were fewer values of 4 and above for the rushed responding than for the patterned responding scale. Only 9 participants were flagged in every scale, 17 in 3 scales, 24 in 2 scales, and a majority of 49 participants were flagged in only 1 of the 4 self-reported scales. In total, the 4 scales flagged 99 (25.1%) participants as conspicuous.

The SRSI UseMe, indicating whether we should use the data provided by the participant or not, was negated by 22 participants (5.6%). Thus, these participants were also flagged as self-reported careless. It was then decided to aggregate these self-reported measures into one

Figure 1. Distributions of self-reported responding tendency scales. A random value was added to individual points to reduce overplotting.



variable for self-reported carelessness.

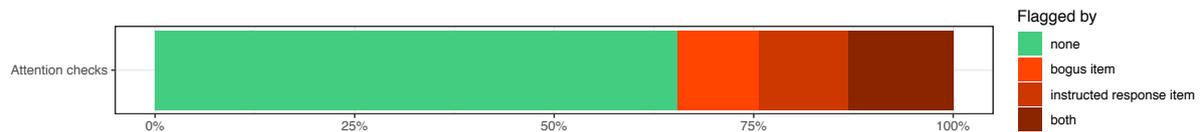
Aggregating the self-reported measures of carelessness, patterned responding, rushed responding (*flagged* ≥ 4) and the SRSI UseMe, 106 participants (26.9%) were flagged as self-reported low-quality responses (see Table 2).

Attention checks. The IRI and the Bogus Item were missed by 96 participants (24.4%) and 92 participants (23.3%), respectively. Because there was no clear cutoff for the Bogus Item, we decided to code all answers with an agreement of 4 or higher to the item "*I see myself as someone who did not read this statement*" as failing to answer the Bogus Item correctly.

Figure 2 demonstrates that the majority of respondents (258, 65.5%) answered both items

correctly, while 40 (10.2%) failed only at the Bogus Item, and 44 (11.2%) only at the IRI. However, a large number of participants who were flagged as inattentive missed both questions (52, 13.2%).

Figure 2. Number of participants flagged by one or both attention check items.



Post hoc detection methods

The boxplots and individual values of each post hoc detection method are presented in Figure 3. Where applicable, cutoffs are indicated by a vertical line and flagged participants are marked red ("fail") and inconspicuous participants are marked blue ("pass").

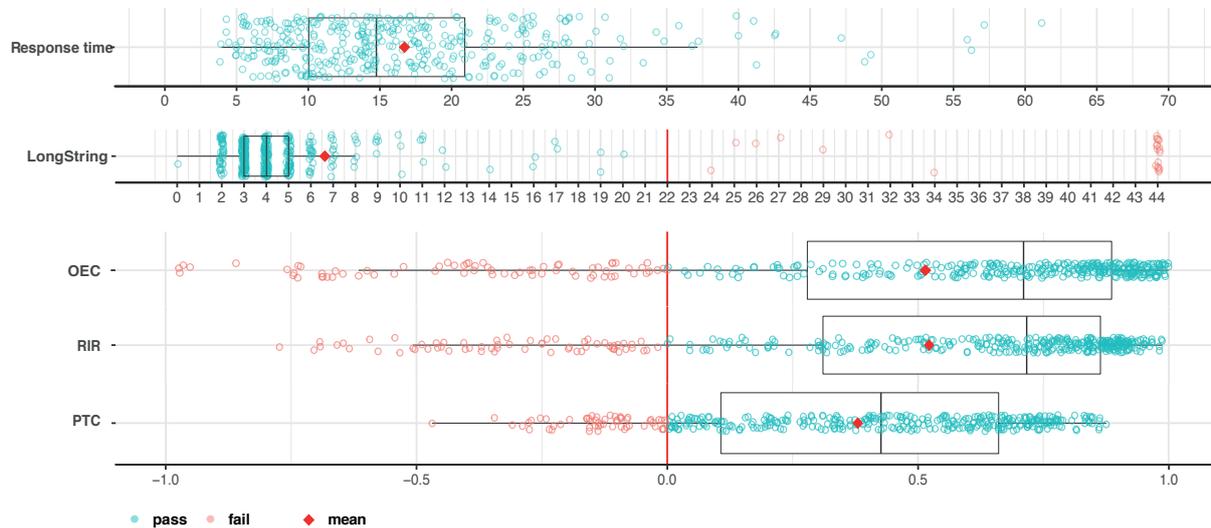
Response Time. Although Huang et al. (2012) recommended a general cutoff for too quick response times (2s an item), the distribution presented in Figure 3 did not show a cluster or conspicuous responses below a certain value. Therefore, no suspiciously fast respondents were flagged.

LongString Index. Results from the LongString analysis, with the recommended cutoff from Curran (2016) (>22), reveal that 22 (6.3%) of the participants were flagged by this method. The distribution depicted in Figure 3 displays that the vast majority of participants were significantly below this threshold, and 18 (4.6%) suspicious respondents with a LongString Index of 44 were identified with this method. These 18 participants provided the same answer for all 44 items of the BFI.

Odd-even consistency. The distribution in Figure 3 is left-skewed with a long tail, and only a few suspicious correlations are close to -1 . Curran (2016) recommended removing all correlations below 0, which in this case would flag 63 (16.0%) of participants as responding too inconsistently.

Resampled individual reliability. As a more general approach to consistency than the OEC, RIR was calculated with 100 times randomly selected two halves of each subscale of the BFI. These two vectors were then correlated for each individual, giving a more general

Figure 3. Boxplots of carelessness detection methods. OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation. Response time in minutes for the entire survey.



(resampled) reliability. As with the OEC, the distribution is left-skewed with a long tail (see Figure 3). However, it has less extreme negative values, and slightly less respondents were identified as careless with this method (61, 15.5%).

Person-total correlation. Correlations of individual answers with the mean of answers from the whole sample exhibited a comparatively narrow distribution of values (see Figure 3). This method flagged 74 (18.8%) of participants as careless, indicated by a correlation of less than 0.

Open answer quality

Open answer quality was coded either 0 (= Insufficient), 1 (= High), or 2 (= Excellent). Of the full sample, 100 participants (25.4%) displayed insufficient answer quality in the open question. As we are mainly interested in whether participants failed or succeeded to provide sufficient open answer quality, the high (146, 37.1%) and excellent (148, 37.6%) open answer quality categories were combined for further analysis.

Relationship between open answer quality and carelessness detection methods. The 100 participants providing insufficient open answer quality will be referred to as the IAQ group in this section. Accordingly, the SAQ group represents the 294 participants with sufficient open answer quality. Results demonstrated that participants with IAQ significantly more often

failed the IRI ($\chi^2(1, N = 394) = 21.35, p < .001$) as well as the Bogus Item ($\chi^2(1, N = 394) = 24.665, p < .001$) than the SAQ group. Furthermore, 43% of all participants with IAQ were flagged by the self-reported carelessness measures, while only 21.4% were flagged in the SAQ group. The LongString cutoff flagged 15% of all IAQ participants and 3.4% of participants with SAQ. The 100 participants with low answer quality displayed higher LongString values ($M = 9.55, SD = 12.42$) than those with high quality ($M = 5.63, SD = 7.49$). A Wilcoxon rank-sum test yielded a significant difference at an alpha level of 5% ($W = 17730, p < .01$). Moreover, 24% of the IAQ group and 13.3% of the SAQ group failed the OEC cutoff. Similarly, 29% in group IAQ and 10.9% of group SAQ failed to display an RIR above the cutoff. For the PTC, 29% of group IAQ and 15.4% of group SAQ failed to display a positive correlation between their answer with the rest of the sample. Lastly, participants in the IAQ group ($M = 882.58$ seconds, $SD = 563.66$ seconds, $n = 97$) needed significantly less time to complete the survey than participants in the SAQ group ($M = 1042.83, SD = 544.93, n = 289$), in a Wilcoxon rank-sum test, $W = 10958, p < .01$.

Correlations between carelessness detection methods

Table 3 depicts how successfully the different methods correlate in their decision to classify participants either as suspicious or not suspicious. Answer quality achieved relatively low correlation with all behavioral and self-report measures of carelessness. The highest correlations of answer quality were observed with the Bogus Item (.26) and the IRI (.24). Interestingly, while the IRI and Bogus Item correlated with .41, the Bogus Item exhibited a higher correlation with the consistency measures PTC (.52), RIR (.51), LongString (.37), and OEC (.36). Self-reported data quality correlated substantially with RIR (.38), the Bogus Item (.34) and the IRI (.30). Unsurprisingly, the highest correlation was observed between OEC and RIR (.68), because RIR is a generalization of OEC. Overall, the correlation pattern demonstrates that among the attention check items the Bogus Item correlated more strongly with several other measures when compared to the IRI. The LongString Index exhibits similar correlations with all behavioral measures, except with IRI. The consistency measures correlate strongly with each other, apart from a relatively weak correlation between OEC and RIR (.25).

However, the relationship of answer quality with other measures is less clear. Based on these correlations, it is difficult to claim that one of the measures is redundant, as all the measures have relatively low overlap.

Table 3

Matthews correlation coefficient (MCC) of each measure pair (N = 394). A value near 1 suggests that the two methods have a high overlap in the classification of careless/not careless participants. IRI = Instructed Response Item, OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation.

	1.	2.	3.	4.	5.	6.	7.
1. Self-report	-						
2. Bogus Item	.34	-					
3. IRI	.30	.41	-				
4. LongString	.24	.37	.26	-			
5. OEC	.23	.36	.20	.40	-		
6. RIR	.38	.51	.22	.37	.68	-	
7. PTC	.31	.52	.27	.35	.25	.41	-
8. Answer quality	.21	.26	.24	.21	.13	.22	.15

Classification of respondents based on different methods.

Latent profile analysis. To identify different classes of carelessness, a Latent profile analysis (LPA) was conducted. Latent profile analysis is a flexible model-based approach to classification, with less restrictive assumptions than cluster analysis (Muthén, 2002). It aims to find the smallest number of profiles that can describe associations among a set of variables, and a formal set of objective criteria are applied to identify the optimal number of latent profiles in the data. For each participant, LPA provides a probability of membership, which is based on the degree of similarity with each prototypical latent profile. Following the approach by Meade and Craig (2012), we conducted an LPA on the non-self-report indicators of response quality (Open Answer quality, Response time, IRI, Bogus Item, LongString Index, OEC, RIR, and PTC) using the *mclust* package for R (Scrucca, Fop, Murphy, & Raftery,

2016). Self-report indicators were excluded, enabling a comparison of our results with Meade and Craig (2012), and because these indicators might be biased when participants are paid to participate. However, the self-report indicators were subsequently used to describe the different classes found in our data.

Open answer quality, IRI, and Bogus Item were binary variables (pass/fail). Missing data was present because for participants with a LongString Index of 44 (all items with the same answer) no OEC, RIR, and PTC measures could be computed (no variance in the answers). We therefore inputted missing values in these variables with +1 for consistency and reliability and -1 for PTC. Missing values in the response time variable were possible if participants did not respond to the questionnaire in one sitting. These missing values were estimated using an expectation maximization algorithm as implemented in mclust. Based on these variables, multiple models with one to nine classes were fitted. Bayesian information criterion (BIC) and integrated complete-data likelihood (ICL) criterion were used to judge the most appropriate number of classes. Both indicated that three classes were most appropriate (BIC: -7404.41, ICL: -7414.34). The class sizes were 181 (45.9%) for class 1, 129 (32.7%) for class 2, and 84 (21.3%) for class 3. The frequencies and variable means associated with each class are presented in Table 4.

As shown in Table 4, answers from class 1 were frequently judged as insufficient quality. Moreover, the attention check items were only missed by participants from this class. Further, class 1 participants more frequently self-reported bad quality than those in classes 2 and 3. Classes 1 and 2 responded significantly more quickly than class 3. Concerning OEC, class 3 provided more inconsistent answers than classes 1 and 2. Additionally, class 3 showed slightly stronger agreement to the self-reported responding tendencies than class 2. The defining hallmarks of class 1 were very large LongString Index values and very low PTC. This demonstrates that the consistency within participant answers was relatively high, while these answers were noticeably different from the total sample. Overall, it appears that a large part of class 1, which accounts for 45.9% of the sample, was responding in a careless way. However, class 1 cannot be described by one singular measure of carelessness. Instead, several forms captured by different methods should be included. In contrast, class 2 displayed the best values

Table 4

Descriptive statistics for each identified class of participants. IRI = Instructed Response Item, OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation.

Variable	Class 1	Class 2	Class 3
Class size	181 (45.9%)	129 (32.7%)	84 (21.3%)
Percentages pass			
Answer quality (%)	44.75	100	100
Bogus Item (%)	49.17	100	100
IRI (%)	46.96	100	100
Self-report (%)	59.12	90.70	76.19
SRSI UseMe (%)	90.06	99.22	96.43
Means			
Response time (in Minutes)	14.58	16.94	22.03
OEC	.52	.86	.37
RIR	.44	.83	.43
PTC	.13	.59	.30
LongString	9.61	3.79	4.83
Means (Self-reported)			
Careless responding	2.16	1.36	1.52
Patterned responding	2.28	1.20	1.49
Rushed responding	2.43	1.68	1.88
Skipping instructions	2.53	1.99	2.13

for all examined measures. Class 3 was slightly more conspicuous in terms of self-reported scales, OEC, RIR, and PTC. This class appeared to answer slightly less consistently than class 2, but still managed to pass all attention checks and to provide sufficient open answer quality.

Prediction of class membership. It might not always be possible to incorporate all the above-mentioned carelessness detection methods in a study. Therefore, it was of interest to reduce the number of measures but still be able to identify participants of the careless class 1 accurately. Conditional inference trees, as implemented in the *party* package for *R* (Hothorn, Hornik, & Zeileis, 2006), were used to identify the most predictive measures for class membership for each participant. Conditional inference trees use a recursive algorithm to make an unbiased selection among covariates, and offer several advantages over traditional regression models and random forests (Hothorn et al., 2006; Strobl, Malley, & Tutz, 2009), such as non-linear relationships and less overfitting. Nine variables were used to predict class membership (SRSI UseMe, Bogus Item, IRI, Response time, LongString, OEC, RIR, PTC, and Open answer quality). The SRSI UseMe variable, Bogus Item, IRI, and the answer quality were included as binary variables (Pass/Fail), whereas the remaining variables were used in their raw form. Results of the analysis depicted in Figure 4 demonstrate that answer quality, IRI, and Bogus Item are well suited to separate the careless class 1 from classes 2 and 3. Furthermore, taking post hoc detection methods such as LongString analysis, OEC, PTC, and Response time into account, the tree successfully separates classes 2 and 3. Table 5 demonstrates that the prediction based on this model is very accurate (Accuracy = .987, 95% CI [.971, .996]) in terms of identifying the correct class membership. Only 5 participants out of 394 were assigned to the wrong class based on this model.

Effects of carelessness on experimental manipulation

The goal of the experiment included in the study was to examine how effect sizes and p-values changed when careless participants were excluded from the analysis. Results of a Welch's t-test with the full sample demonstrated that there was a significant difference in perceived trustworthiness of the online shop, $t(381.83) = 5.64$, $p = 3.344e - 08$, $d = 0.567$. Participants who saw the low-trust website mock-up rated the company as less trustworthy ($M = 4.36$, SD

Figure 4. Conditional inference tree for all carelessness detection methods. For each inner node, the Bonferroni-adjusted p-values are presented, the fraction of participants in each class (1, 2, or 3) is displayed for every terminal node.

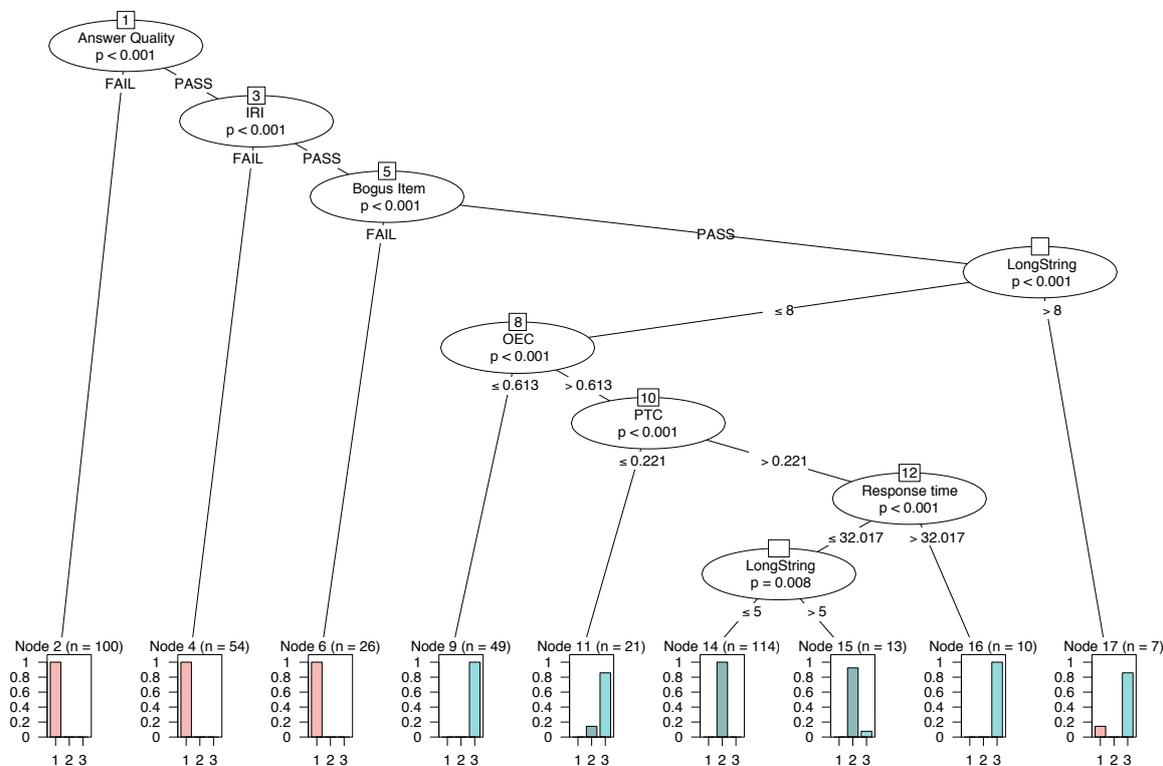


Table 5

Performance of the conditional inference tree model in predicting class membership.

	Class 1	Class 2	Class 3
Predicted			
Class 1	180	0	0
Class 2	0	126	1
Class 3	1	3	83

Note. N = 394

= 1.21) than participants in the high-trust condition (M = 4.99, SD = 1). When participants from class 1 (n = 181) were removed, participants in the low-trust condition rated the company slightly less trustworthy (M = 4.34, SD = 1.19), and participants in the high-trust condition rated the website somewhat more trustworthy than the full sample (M = 5.12, SD =

0.95). The standard deviations decreased slightly in both groups, which indicates that some of the noise that could stem from careless participants was reduced. Although the differences between the two conditions was significant in both cases, removing participants from the careless class 1 led to a smaller p-value and increased the effect size, $t(194.32) = 5.83$, $p = 2.277e - 08$, $d = 0.803$.

Discussion

Analysis of careless behavior in a crowdsourced sample

Previous work studied carelessness or other deceptive forms of behavior in online samples either with only a few methods (J. J. Chandler & Paolacci, 2017; Hauser & Schwarz, 2016; Kan & Drummey, 2018; Peer et al., 2017), or they assessed carelessness in student samples or mixed online samples (Maniaci & Rogge, 2014; Meade & Craig, 2012). We build on this work with a systematic analysis of carelessness in a crowdsourced sample and examine the new methods: RIR and PTC (Curran, 2016). We applied six measures and corresponding cutoffs, based on recommendations from Meade and Craig (2012), Maniaci and Rogge (2014), and Curran (2016), to identify multiple forms of carelessness in a crowdsourced sample from FigureEight.

Observing the planned detection methods, which require special items or scales, 26.9% of all participants indicated in self-reports to provide careless, patterned, or rushed responses, 24.4% of all participants failed to answer the IRI correctly, and 23.4% missed the Bogus Item (see Table 2). Weak to moderate correlations between aggregated self-reported carelessness and other detection methods only partially indicate convergent validity for self-report measures (see Table 3). Correlations between attention check items and other detection methods were also weak to moderate, except for the Bogus Item that correlated relatively strongly with RIR and PTC. The 24.4% of participants in our FigureEight sample who failed the IRI surpass the 14% found in the study by Maniaci and Rogge (2014), which examined a sample including MTurk workers, participants from online forums, and psychology students. This indicates that inattentive behavior may be more frequent in samples from crowdsourcing platforms. Taken together, approximately 25% of the sample was flagged as inattentive, based solely on one of

the attention check items. It can be expected that the overall number of participants flagged with these items increases with the length of the survey, as one attention check item is recommended for every 50–100 items (Meade & Craig, 2012). In a considerably longer study, applying 4 attention check items, Peer et al. (2017) found 73% of all participants fail at least one attention check item. Post hoc detection methods revealed 6.3% of all participants were flagged by the LongString analysis, which corresponds with findings from Maniaci and Rogge (2014), where 6% were flagged in a mixed sample. However, the OEC and RIR revealed 16% and 15.5%, respectively, as responding too inconsistently. This is more than twice as much as Maniaci and Rogge (2014) identified with the OEC method. Lastly, the PTC flagged 18.8% of all participants as being careless. Hence, the post hoc detection methods of the present study further suggest that careless or inattentive behavior may be more pronounced in a fully crowdsourced sample.

How prevalent is careless responding in samples from crowdsourcing platforms, based on various detection methods for carelessness? (RQ1)

Almost 60% of all participants were flagged by at least one of the methods examined in this study (see Table 2). However, the univariate examination of single measures, and a subsequent cumulative exclusion of participants, might be problematic for various reasons. Firstly, with this strategy, participants are excluded based on methods that do not have a set cutoff or an objective wrong answer, and the researcher has to decide whether one or multiple flags per participant would lead to an exclusion from the sample. Secondly, simply combining the different measures altogether might be too restrictive and lead to many false-positives. For instance, the PTC cutoff might not be meaningful in situations where a lot of carelessness can be expected. Therefore, and in line with Maniaci and Rogge (2014) and Meade and Craig (2012), we base our prevalence estimate of carelessness on the results of the LPA, which takes multiple raw values of various non-self-report methods into account to identify different classes of participants.

The LPA identified three classes in total. Class 1 (the careless participant class), contained 45.9% of all participants. Although this class cannot be described by one measure, and

therefore comprises multiple forms of inattention and carelessness, its characteristics can be summarized as follows: Failing in providing sufficient open answer quality and failing attention checks was an exclusive characteristic of this class. This class also self-indicated bad data quality considerably more often than the other two classes. Participants of this class answered more quickly, showed very large LongString values, and a very low PTC, indicating excessive consistency within, yet low congruence with the total sample. While the OEC measure also revealed a relatively high inconsistency in the answers of this class, it is important to note that class 3, usually inconspicuous concerning other detection methods of carelessness, provided even more inconsistent answers. This finding suggests using caution with measures of consistency as a means of data cleaning, because they might bear potential for a high false-positive rate. The LPA from the present study revealed a considerably larger group of careless participants (45.9%) in a crowdsourced sample compared to similar analyses conducted with mixed online samples or student pools in the studies in Maniaci and Rogge (2014) and Meade and Craig (2012). These studies identified approximately 2.2% to 11% as being careless. Concerns surrounding the representativeness of a sample after excluding such a large percentage of participants, and from an economic perspective, might suggest not using such a sample.

How are task-specific measures related to planned detection methods and post hoc methods? (RQ2)

Out of 394 participants, 100 (25.4%) provided insufficient open answer quality. Significantly fewer participants of this group passed attention checks; they more often self-reported bad data quality and they exhibited significantly higher LongString Index values. Furthermore, participants who failed in providing sufficient answer quality completed the survey in significantly less time, they more often failed to meet the OEC cutoff and the RIR, and they more often failed to meet the PTC cutoff. Hence, these results indicate some convergent validity for open answer quality as a measurement for carelessness. However, correlations between this measure and other planned detection or post hoc methods were rather weak (see Table 3). This might indicate that carelessness depends (to a large extent) on the given task.

This result coincides with findings from Maniaci and Rogge (2014), indicating that inattention or carelessness during specific tasks (such as watching a video or marking pronouns in a text), mostly has a low correlation with other detection methods of carelessness. Therefore, completing standardized Likert-scale questionnaires, answering open questions, watching videos, and participating in concentration tasks in online studies appears to evoke different forms of inattention, which tend to concern different participants.

Which methods are most applicable to identify carelessness in a crowdsourced sample can be made, based on our findings? (RQ3)

In general, we strongly encourage other researchers to analyze the data quality of crowdsourced surveys. As in Maniaci and Rogge (2014) and Meade and Craig (2012), we refer to the LPA as our reference for careless behavior in our sample. Based on our findings, we recommend a set of measures that are easy to apply, easy to interpret, and at the same time cover the majority of the inattentive class 1.

1. Further, we recommend including an SRSI UseMe item to assess whether participants indicate that their data should be used for the study. Although this item was not an important predictor of class membership, it acts as a form of revoked consent. Thus, it serves a purpose beyond detecting bad data quality. However, from a practical perspective, we cannot currently recommend other self-report measures. This is because including 10 or more additional items in a survey with the sole purpose of detecting self-reported bad data quality may not be an efficient approach for all online surveys.
2. Attention checks such as an IRI should be included, because these detection methods are easy to create and offer a clear interpretation. We further advise to include a Bogus Item, as the combination of a task-specific measure, the IRI, and the Bogus Item was successful in classifying 180 of 181 participants correctly in class 1. However, the wording of the Bogus Item should be chosen carefully, because Bogus Items can cause interpretative problems (Meade & Craig, 2012).
3. The inclusion of the LongString Index as a post hoc measure is recommended, because this measure is applicable to all type of scales (given sufficient length), and provides an

overview of repetitive answer patterns. Moreover, a high LongString Index was a typical characteristic of the inattentive class 1 (see Table 4). However, in most cases, a task-specific measure and a combination of attention checks appears sufficient, as the LongString Index was not a significant predictor for class 1 in the conditional inference tree. Based on the results of the conditional inference tree and the LPA, we cannot recommend the other post hoc detection methods. The minor response time differences between classes 1 and 2 did not offer a clear and readily applicable cutoff value for an inattentive class (see Table 4). Furthermore, response time was not identified as a significant predictor of class 1. Concerning OEC, the LPA identified a class (class 3, see Table 4), with slightly lower values than the careless class 1. However, apart from this measure, and a considerably longer average response time, this class was inconspicuous. Thus, flagging participants based on this measure might lead to a high false-positive rate. The RIR of class 1 was comparable to class 3. However, although class 1 showed a lower PTC (see Table 4), this measure was not predictive for class 1 in the model. In comparison to a LongString analysis, the interpretation of this measure is more dependent on the properties of a sample. In a sample with poor data quality, this measure is heavily biased, because it correlates individual responses with averaged responses, including all careless participants (Curran, 2016).

4. Results suggest that carelessness is dependent on the given tasks to participants. The correlation table (see Table 3) demonstrates that the open answer quality achieved relatively low correspondence with other detection methods for carelessness. Meanwhile, low open answer quality is exclusively (and thus clearly) associated with the inattentive class 1. Therefore, we encourage researchers to apply carelessness detection methods according to the given tasks. While planned and post hoc detection methods might generally identify carelessness in Likert-type questionnaires, they might not prevent bad data quality in other types of online-survey tasks.

Taken together, we recommend the following set of carelessness detection methods: an SRSI UseMe item, one or multiple Instructed Response or Bogus Items, a LongString analysis, and a task-specific measure (in our case: assessing open answer quality). These measures either

represented important predictors for the inattentive class 1, or they provided pragmatic merit for analyzing the data quality of a crowdsourced sample. All these methods are relatively straightforward to apply, as they do not need to consider scale dimensions and inverse items. Furthermore, they were clearly associated with the inattentive class 1 in the LPA, and the prediction for class 1 (based on these detection methods) was very accurate: 180 out of 181 were correctly identified, while none of the participants from classes 2 and 3 were falsely flagged by this combination of methods. As demonstrated by the experiment included in this study, removing careless participants increased effect sizes from $d = 0.567$ to $d = 0.803$. Although the difference was very robust in the sample including careless respondents, research has also shown that carelessness can not only reduce effects but also disperse known effects (e.g., DeSimone & Harms, 2018; Maniaci & Rogge, 2014). Hence, carelessness may reduce statistical power and increase noise in the data, thus undermining the validity of online experiments. Therefore, it is vital that researchers develop a data cleaning strategy whenever online samples are recruited, and cleaning process must be reported in detail.

Limitations and future research

Some limitations must be considered concerning the results and recommendations presented in this paper:

First, the present study was conducted on the FigureEight platform, and this might not readily translate to other platforms or recruitment methods. For instance, Amazon's MTurk offers different methods of community-management and rating possibilities for workers, which may cause workers to be more attentive when taking part in a survey. Future research, therefore, should systematically assess data quality differences between various platforms, applying multiple carelessness detection methods.

Second, the present study assessed the detection of careless participants, which resulted in the exclusion of approximately half of all participants. Excluding this number of participants could have severe methodological and financial implications. Hence, future research should also focus on preventing carelessness, which is presently not well understood. Warnings about monitoring data quality that have been used by Clifford and Jerit (2015) or Meade and Craig

(2012) can be effective, but might lead to other biases, such as socially desirable behavior.

D. Chandler and Kapelner (2013) have found positive effects of meaning by explaining the purpose of a task on data quality in crowdsourcing tasks. Furthermore, Ward and Pond (2015) found that promising participants the results of the study was effective in increasing data quality. More effort is needed to systematically analyze these measures for preventing carelessness in crowdsourced samples.

Third, the present study included an open-ended question to assess task-dependency of careless behavior. Although findings from Maniaci and Rogge (2014) suggested a similar approach (by applying other forms of different tasks in their survey), future research should aim for a systematic review of a wider variety of different tasks in online surveys. This will facilitate further analysis of the task-dependency of careless behavior.

Finally, as all post hoc detection methods are approximate and uncertain, bad data quality can not clearly and reliably be identified in every case. Our recommendations are based on the prediction of class 1, which was identified using LPA. Only planned detection methods were found to be predictive for this class. However, there are situations where it might not be possible to include attention check items or task-dependent measures of quality, such as voluntary surveys of highly specific populations. Hence, further research is needed to ensure data quality in such situations.

Conclusion

The aim of this study was to provide an estimate of the frequency of carelessness in samples from crowdsourcing platforms, based on different identification methods. Our results reveal that approximately half of all crowdsourced participants display careless behavior.

Furthermore, carelessness and inattention appear highly task-dependent, as correlations between open answer quality and other measures are rather low. Finally, based on a predictive model and interpretative problems of several detection methods, we recommend assessing data quality of crowdsourced samples by applying the following: an SRSI UseMe item, attention checks such as the IRI and Bogus Item, the LongString Index, and a task-specific measure. A combination of these methods was able to identify 180 out of 181 inattentive

participants, and the subsequent exclusion of this subsample resulted in an increased effect size and smaller p-value in the experiment.

Acknowledgments

This work has been approved by the Institutional Review Board of the Faculty of Psychology, University of Basel under the number D-006-17. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*(2), 527–535. doi: 10.3758/s13428-012-0265-2
- Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160. doi: 10.1016/j.chb.2013.05.009
- Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization, 90*, 123–133. doi: 10.1016/j.jebo.2013.03.003
- Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130. doi: 10.3758/s13428-013-0365-7
- Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science, 26*(7), 1131–1139. doi: 10.1177/0956797615585115
- Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*(1), 53–81. doi: 10.1146/annurev-clinpsy-021815-093623
- Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science, 8*(5), 500–508. doi: 10.1177/1948550617698203
- Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon mechanical turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology, 32*(4), 347–361. doi: 10.1007/s10869-016-9458-5
- Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly, 79*(3), 790–802. doi: 10.1093/poq/nfv027
- Comley, P. (2015). Online market research. In M. v. Hamersveld & C. d. Bont (Eds.), *Market*

- Research Handbook* (pp. 401–419). Chichester, England: John Wiley & Sons Ltd. doi: 10.1002/9781119208044.ch21
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. doi: 10.1016/j.jesp.2015.07.006
- DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, 33(5), 559–577. doi: 10.1007/s10869-017-9514-9
- de Winter, J., Kyriakidis, M., Dodou, D., & Happee, R. (2015). Using crowdflower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing*, 3, 2518–2525. doi: 10.1016/j.promfg.2015.07.514
- Dogan, V. (2018). A novel method for detecting careless respondents in survey data: floodlight detection of careless respondents. *Journal of Marketing Analytics*, 6(3), 95–104. doi: 10.1057/s41270-018-0035-9
- Douglas, K. M., & McGarty, C. (2001). Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology*, 40(3), 399–416. doi: 10.1348/014466601164894
- Flavián, C., Guinalú, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, 43(1), 1–14. doi: 10.1016/j.im.2005.01.002
- Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in mturk and other online samples. *Industrial and Organizational Psychology*, 8(2), 196–202. doi: 10.1017/iop.2015.25
- Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1631–1640). New York, NY, USA: ACM. doi: 10.1145/2702123.2702443
- Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, 66(1), 877–902. doi: 10.1146/annurev-psych-010814-015321
- Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung

- wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Wiesbaden: Vieweg+Teubner Verlag. doi: 10.1007/978-3-322-80058-9_19
- Hauser, D. J., & Schwarz, N. (2016, Mar 01). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1), 400–407. doi: 10.3758/s13428-015-0578-z
- Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review*, 27(2), 196-212. doi: 10.1177/0894439308327481
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674. doi: 10.1198/106186006X133933
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. doi: 10.1007/s10869-011-9231-8
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828. doi: 10.1037/a0038510
- John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (Vol. 2, pp. 102–138). New York, NY, US: Guilford Press.
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39(1), 103–129. doi: 10.1016/j.jrp.2004.09.009
- Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, 18(3), 512–541. doi: 10.1177/1094428115571894
- Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? an examination of worker misrepresentation and downstream consequences in amazon's mechanical turk

- workforce. *Computers in Human Behavior*, 83, 243–253. doi: 10.1016/j.chb.2018.02.005
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. doi: 10.1016/j.jcm.2016.02.012
- Lee, H. (2006). Privacy, publicity, and accountability of self-presentation in an on-line discussion group. *Sociological Inquiry*, 76(1), 1–22. doi: 10.1111/j.1475-682X.2006.00142.x
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83. doi: 10.1016/j.jrp.2013.09.008
- McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior*, 84, 295 - 303. doi: 10.1016/j.chb.2018.03.007
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. doi: 10.1037/a0028085
- Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, 68(10), 689–709. doi: 10.1016/j.ijhcs.2010.05.006
- Muthén, B. O. (2002). Beyond sem: General latent variable modeling. *Behaviormetrika*, 29(1), 81–117. doi: 10.2333/bhmk.29.81
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63, 1–11. doi: 10.1016/j.jrp.2016.04.010
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872. doi: 10.1016/j.jesp.2009.03.009
- Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184–188. doi:

10.1177/0963721414531598

- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. doi: 10.1016/j.jesp.2017.01.006
- Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 205–233.
- 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. doi: 10.1016/j.chb.2014.11.064
- Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology, 80*(2), 325. doi: 10.1037/0022-3514.80.2.325
- Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology, 57*(1), 529-555. doi: 10.1146/annurev.psych.57.102904.190048
- Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly, 73*(2), 325–337. doi: 10.1093/poq/nfp029
- Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21*(10), 736 - 748. doi: 10.1016/j.tics.2017.06.007
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348. doi: 10.1037/a0016973
- Tuch, A. N., Schaik, P. V., & Hornbæk, K. (2016). Leisure and work, good and bad: The role of activity domain and valence in modeling user experience. *ACM Transactions on Computer-Human Interaction, 23*(6), 35:1–35:32. doi: 10.1145/2994147
- Van Pelt, C., & Sorokin, A. (2012). Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of*

- Data* (pp. 765–766). New York, NY, USA: ACM. doi: 10.1145/2213836.2213951
- Ward, M., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417 – 430. doi: <https://doi.org/10.1016/j.chb.2017.06.032>
- Ward, M., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554–568. doi: 10.1016/j.chb.2015.01.070
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063. doi: 10.1037/0022-3514.54.6.1063

Appendix

Calculating the open answer quality Index

A-priori criteria for the rating of open-ended questions were defined according to the measures used in studies from Holland and Christian (2009); Smyth et al. (2009). The following indicators for calculating an open answer quality index were taken into consideration:

Substantive response. This indicator refers to whether the participant answer thematically corresponds to the open question subject matter. The open answer has been coded with 0 if it merely consisted of meaningless sequences of letters, clearly copy-pasted phrases, or thematically unfit answers which typically emerged from not carefully reading the instructions (such as describing a negative experience in a non-virtual store instead of an online shop). If the open answer corresponded to the subject matter, regardless whether the participant addressed all subquestions, the indicator has been coded with 1.

Number of words. Because it is possible to provide a thematically substantial answer while providing little or zero actual content (such as merely writing one short sentence about the experience), the number of words has been assessed for each open answer. Given the topic of the open question and the two subsequent subquestions, a minimum of 50 words (± 3) was defined as the requirement to answer the questions. Thus, participants were explicitly asked to

provide an answer containing at least 50 words. This number is regarded as being a minimum effort to achieve a thematically substantial answer that additionally addresses at least one subquestion. Wordcounts corresponding to this number (or higher) were coded with 1, smaller wordcounts with 0.

Complete sentences. Participants were explicitly asked to provide answers with full sentences. Open answers that mainly or exclusively consisted of unfinished sentences (or separate words) were coded with 0 in regard to complete sentences. To receive a coding of 1, the majority of all sentences in the open answer needed to be complete and separated with commas or periods.

Number of subquestions answered. If none of the specific subquestions were addressed, the answer was coded with 0 in regard to number of subquestions. This was also the case if the given answer met the requirements for a thematically substantial answer, but failed to answer at least one of the specific subquestions. Accordingly, the answer received a coding of 1 or 2 if one or both subquestions were addressed in the open answer, respectively.

Number of subquestions elaborated. An answer to a subquestion was considered to be elaborate if the according part of the open answer contained at least three complete sentences. If none of the subquestions were elaborated, the answer was coded with 0 in regard to themes elaborated. Accordingly, the answer received a coding of 1 or 2 if one or both subquestions were elaborated in the open answer, respectively.

Calculation of the open answer quality Index. *Substantive response* and *Number of words* were seen as essential for providing a valuable open answer. Thus, if one or both of these variables were coded with 0, the open answer quality Index was also automatically coded with 0. The other variables, namely *complete sentences*, *number of subquestions answered* and *number of subquestions elaborated*, were seen as being important (but not an absolute necessity) on their own in order to provide a good open answer quality. Thus, for answers that met the minimum requirements, the codings from *complete sentences*, *number of themes*, and *themes elaborated* were counted together. If the sum reached 3 or higher, the overall open answer quality was considered to be adequate, and thus coded with 1.