Unaufmerksamkeit, Faking, Speedster… Kontrolle der Datenqualität in User Experience Befragungen

Meinald T. Thielsch
Department of Psychology
University of Münster
Münster, Germany
thielsch@uni-muenster.de

Gerrit Hirschfeld
Faculty of Business
Bielefeld University of Applied Sciences
Bielefeld, Germany
gerrit.hirschfeld@fh-bielefeld.de

ABSTRACT

User Experience wird oftmals anhand von Befragungen mittels standardisierter Fragebögen erfasst. Allerdings bringen dabei gerade quantitative Online-Datenerhebungen den großen Nachteil mit sich, dass die Umstände, unter denen die Befragten an der Untersuchung teilnehmen zumeist nur schwer zu kontrollieren sind. Insbesondere ist unklar, unter welchen Bedingungen und wie motiviert die Teilnehmer*innen geantwortet haben. Zudem kommen Drop-out und hohe Abbruchquoten in Online-Befragungen durchaus oft vor – was für die Interpretation der Ergebnisse vor allem dann problematisch ist, wenn dieser Non-response selektiv erfolgt.

Kritischer sind Personen, die eine Befragung formal abschließen, aber unbewusst (z.B. aufgrund von Unaufmerksamkeit) oder bewusst (Faking) falsche Angaben machen oder sich wahllos durch einen User Experience Fragebogen „durchklicken“ (sogenannte Speedster). Solche Falschangaben können zu einer fehlerhaften Interpretation der Ergebnisse führen. Während Speedster relativ leicht anhand der Antwortzeit zu identifizieren sind, fällt dies bei Unaufmerksamkeit oder bewusstem Faking schwerer. In unserem Beitrag geben wir hier eine Übersicht über eine Reihe von möglichen Maßnahmen, angefangen bei spezifischen Datenprüfungen über Kontrollfragen bis hin zu gezielten technischen Kontrollen. Diese Maßnahmen werden anhand eines Fallbeispiels einer Website-Evaluation mit 30 Testwebsites und mehreren tausend Befragten illustriert.

CCS CONCEPTS

• Human-centered computing → Human computer interaction (HCI) → Empirical studies in HCI

Mensch und Computer 2021, Workshopband, Workshop on Quantitative Messung von User Experience.

https://doi.org/10.18420/muc2021-mci-ws01-369

KEYWORDS

Datenqualität, Qualitätssicherung, Befragungen, Online-Studien, Durchklicker, bewusste Falschangaben, User Experience Evaluation

ACM Reference format:

Meinald T. Thielsch and Gerrit Hirschfeld. 2021. Unaufmerksamkeit, Faking, Speedster… Kontrolle der Datenqualität in User Experience Befragungen. In *Mensch und Computer 2021 (MuC'21), September 5–8, 2020, Ingolstadt, Germany. ACM, New York, NY, USA, 4 pages.* https://doi.org/10.1145/1234567890

1 Einführung

Bei User Experience geht es im Kern um das Erleben der Nutzer*innen – somit ist es zentral diese zu befragen. Dabei ist den Interviewer*innen meist sehr bewusst, dass es Antworttendenzen und Urteilsfehler gibt. Insbesondere in quantitativen Befragungen können entsprechende Gegenmaßnahmen erfolgen[1]: So werden beispielsweise mehrere Fragen zu Skalen zusammengefasst, um Messfehler zu reduzieren, die Fragen innerhalb dieser Skalen randomisiert dargeboten, um Positionseffekte zu vermeiden, oder Fragen mit verschiedener Antwortrichtung gemischt damit keine Ja-Sage-Tendenz (Akquieszenz) provoziert wird.

Aber nicht nur Antworttendenzen und Urteilsfehler können die Qualität von erhobenen Daten beeinträchtigen, auch Unaufmerksamkeit oder bewusste Falschangaben (Faking) der Befragten können eine gewichtige Rolle spielen. In qualitativen Interviews besteht hier zumindest die Möglichkeit bei inkonsistenten oder fehlerhaften Angaben gezielt nachzufragen. In quantitativen Befragungen ist dies nicht möglich. Insbesondere quantitative Online-Datenerhebungen bringen zudem den großen Nachteil mit sich, dass die Umstände, unter denen die Befragten an der Untersuchung teilnehmen nicht der Kontrolle der Studienleitung unterliegen. Die Identität der Befragten kann somit nicht endgültig geklärt werden und es ist

[1] Für praktische Hinweise siehe https://www.gesis.org/gesis-survey-guidelines/

unklar unter welchen Bedingungen die Teilnehmer*innen geantwortet haben. Die Befragten sind leichter ablenkbar – und im Vergleich zur Laborerhebung ist es für sie einfacher zu täuschen oder unerlaubte Hilfsmittel zu verwenden ([2]; [8]). Gerade wenn man in Online-Studien Such-, Wissens- oder Erinnerungsaufgaben gibt, könnten manche Befragten versucht sein diese nicht wie instruiert, sondern beispielsweise mithilfe einer Suchmaschine schnell zu lösen (insbesondere, wenn es leistungsbezogene Incentives gibt). Zudem nicht ganz auszuschließen sind technische Schwierigkeiten bei einzelnen Befragten, vor allem wenn diese veraltete Browser-Software und ältere oder spezielle Anzeigegeräte nutzen.

Nach unserer Erfahrung sind diese Probleme oftmals den Verantwortlichen bewusst – bei der Kontrolle der Daten aus Online-Befragungen liegt der Fokus zunächst jedoch meist auf die Abbruchquoten (so genannter Drop-out). Diese sind oftmals hoch und es ist auf den ersten Blick erschreckend, wenn nur die Hälfte der Personen die eine Befragung beginnt auch bis zum Ende vollständig antwortet. Drop-out stellt aber vor allem dann ein Problem dar, wenn der Abbruch selektiv erfolgt (vgl. [12]; [17]). Oftmals sehen sich Personen nur die Startseite einer Befragung an und verlassen diese dann direkt wieder, ohne begonnen zu haben. Wenn Personen direkt zu Beginn eine Umfrage abbrechen, muss dies aber nicht zwingend mit der User Experience des Testobjekts zu tun haben – der Abbruch kann auch schlicht in den zeitlichen Anforderungen der Studie oder einer gefühlten Nicht-Passung zur Zielgruppe begründet sein.

Des Weiteren ist es wichtig zu betrachten, welche Form von Non-Response vorliegt (bspw. vollständiger Abbruch vs. das Auslassen einzelner Fragen; vgl. [1]) und wo im Fragebogen ein kompletter Abbruch stattgefunden hat: Der oben beschriebene Abbruch auf der Start- oder ersten Instruktionsseite einer Studie stellt einen eher unproblematischen Drop-out dar. Kritischer sind die Personen, die im Verlauf der Befragung aussteigen. Bei solchen Fällen sind Analysen notwendig, um mögliche inhaltlich bedingte Abbruchursachen zu identifizieren und einen selektiven Abbruch auszuschließen. Ein selektiver Abbruch könnte zum Beispiel durch das Testobjekt einer Befragung, dessen User Experience oder auch persönliche Eigenschaften der Befragten (z.B. mangelndes technisches Wissen, mangelndes Vertrauen in Technik o.ä.) begründet sein.

Neben der Frage des Drop-Outs ist aber zentral, welche Qualität die gegebenen Antworten haben. Hier kann es für die Datenqualität und damit für die Validität der abgeleiteten Interpretationen fatal sein, wenn Testpersonen Aufgaben nicht mit der notwendigen Konzentration durchführen oder – schlimmer noch – bewusst falsche Angaben machen oder sich nur durch eine Studie durchklicken, um möglichst schnell ein versprochenes Incentive zu erhalten (sogenannte „Speedster"). Hier haben Online-Befragungen den großen Vorteil, dass es möglich ist das Antwortverhalten der Befragten im Sinne einer Qualitätssicherung sinnvoll zu analysieren. Dies, sowie das mögliche Ausmaß von Problemen wie Unaufmerksamkeit, Faking und Speedster, soll das nachfolgende Beispiel veranschaulichen.

2 Praxisbeispiel: Website-Evaluationsstudie

Im Rahmen einer von der Bundeszentrale für gesundheitliche Aufklärung (BZgA) geförderten Studie sollten insgesamt 30 Websites aus dem Gesundheitsbereich evaluiert werden (Ergebnisse finden sich u.a. in [6] und [16]). Im Fokus waren unter anderem die Konstrukte Inhalt, Usability, Ästhetik und Vertrauen. Es wurde jeder befragten Person zufällig eine der 30 Websites zur Bewertung zugeteilt. Die Erhebung erfolgte über einen kommerziellen Felddienstleister, jede*r Befragte erhielt 1,25 € und ein allgemeines inhaltliches Feedback für die vollständige Studienteilnahme. Ziel war eine in den Merkmalen Alter, Geschlecht und Bildungsgrad für Deutschland bevölkerungsrepräsentative Stichprobe > 2000 Personen, sodass jede Website von mehr als 50 Personen bewertet werden würde.

Zur Erreichung dieses Ziels wurden ca. 15.000 Personen entsprechend den Schichtungskriterien zur Studie eingeladen. Insgesamt 5.020 Personen öffneten die Umfrage, allerdings brachen 1.303 von diesen die Studie ab (529 bereits auf den ersten beiden Befragungsseiten) und 79 widersprachen am Ende der Umfrage der Datenverwendung. Liegen nun 3.638 Datensätze vor mit denen die Ergebnisse der Studie berechnet werden könnten? Nein. Von diesen 3.638 Datensätzen erwiesen sich nur gut ein Drittel (35,3 %, n = 1.284) im Rahmen der Datenkontrollen als vollständig unauffällig. Bei den übrigen Daten fand sich mindestens eine, oftmals mehrere Auffälligkeiten: Von den 3.638 Befragten...

- ... waren 19,6 % unrealistisch schnell (Antwortzeit < 8 Minuten, im Vergleich dazu: die mittlere Antwortzeit aller Befragten lag bei > 20 Minuten für die insgesamt mehr als 150 Fragen + Nutzung der Testwebsite),
- ... waren weitere 10,1 % auffällig schnell (Antwortzeit < 12 Minuten),
- ... absolvierten 28,9 % offensichtlich nicht die Testaufgabe, das Öffnen der zugewiesenen Testwebsite (dies wurde technisch durch ein JavaScript geprüft),
- ... fakten 25,9 % bei den Erinnerungsfragen zur Website und öffneten nun die Testwebsite/eine andere Website (obwohl instruiert wurde in der Befragung zu bleiben),
- ... beantworteten 8,1 % eine erste Kontrollfrage zur Qualitätssicherung falsch (die Itemformulierung war: „Aus technischen Gründen bitte bei dieser Frage mit stimme zu antworten."),
- ... beantworteten 10,2 % die zweite Kontrollfrage zur Qualitätssicherung falsch (die Itemformulierung war: „Zur Qualitätsprüfung hier bitte mit zwei antworten."),
- ... gaben 0,7 % an, das Internet seit 50 Jahren oder länger zu nutzen,
- ... gaben 0,4 % an mehr als 20 h pro Tag *aktiv* online zu sein.

In der nachfolgenden Datenbereinigung wurden diese verschiedenen Indikatoren kombiniert berücksichtigt. Manche waren K.O.-Kriterien (z.B. unrealistisch schnelle Antwortzeit, Nichterfüllen der Testaufgabe), bei anderen erfolgte ein Ausschluss nur wenn mehrere Indikatoren auffällig waren (z.B. wurde das Falschbeantworten einer Kontrollfrage toleriert, wenn die zweite Kontrollfrage richtig beantwortet wurde). Am Ende mussten vor der Auswertung 37,8 % der Daten (insgesamt n = 1.373) aufgrund offensichtlicher Falschangaben und Mängel ausgeschlossen werden, nur 2.265 Personen gingen in die finale Auswertung ein. Diese waren im Durchschnitt 51,8 Jahre alt – und damit älter als die Bundesbevölkerung. An diesem Punkt konnte die angestrebte Repräsentativität nach Ausschluss mangelhafter Daten nicht mehr vollständig erreicht werden. Zudem wurden in die nachfolgenden Auswertungen 981 Datensätze eingeschlossen, bei denen ein Indikator auffällig war. Auch wenn diese jeweils gefundene Auffälligkeit nicht so gravierend war, dass ein Ausschluss der Daten zwingend notwendig erschien, es bleibt doch ein gewisses Restrisiko, dass dieser Anteil der Daten nicht optimal zur Beantwortung der Fragestellungen geeignet sein könnte.

Abschließend zeigt ein Blick in die Daten selbst, dass ohne eine Qualitätskontrolle die Ergebnisse in unserem Praxisbeispiel verzerrt worden wären – und zwar stets in Richtung der Mitte der jeweiligen Bewertungsskalen. Abbildung 1 stellt die Datensätze, die ausgeschlossen wurden (und bei denen Website-Evaluationen vorlagen), denen gegenüber, die auch nach dem Qualitätscheck in der Analyse belassen wurden. Dargestellt sind hier Ergebnisse zu verschiedenen User Experience Indikatoren:

- Usability, erfasst mittels der PWU-Skala ([5]; deutsche Version siehe [13]) und der UMUX-Lite [7]
- Weiterempfehlungsintention (in Abbildung 1: „Recomm.“) und Wiederbesuchsintention (in Abbildung 1: „Revisit“; Items zu beiden Aspekten auf Basis von Moshagen & Thielsch [10])
- Vertrauen in den Anbieter der Website („Trust“; eigene deutsche Übersetzung von McKnight et al. [9])
- Ästhetik, erfasst mittels des VisAWI-S [11]
- Inhaltswahrnehmung, erfasst mittels des Web-CLIC [14]
- Gesamtbewertung der Website (in Abbildung 1: „Webnote“) auf einer Schulnotenskala von sehr gut (1) bis ungenügend (6)

In den ausgeschlossenen Daten zeigen sich im Mittel auf allen Indikatoren durchweg leicht schlechtere Bewertungen (alle Unterschiede sind signifikant). Besonders betroffen davon sind die beiden eingesetzten Usability-Skalen (PWU und UMUX-Lite). Ein möglicher Grund könnte darin liegen, dass unaufmerksame Personen, Faker oder Speedster anscheinend eher im Bereich der Skalenmitte klicken. Die gefundenen Unterschiede sind nicht riesig, würden aber in diesem Praxisbeispiel ohne Kontrolle der Daten zu einer systematischen Unterschätzung der User Experience Indikatoren der getesteten Websites führen.

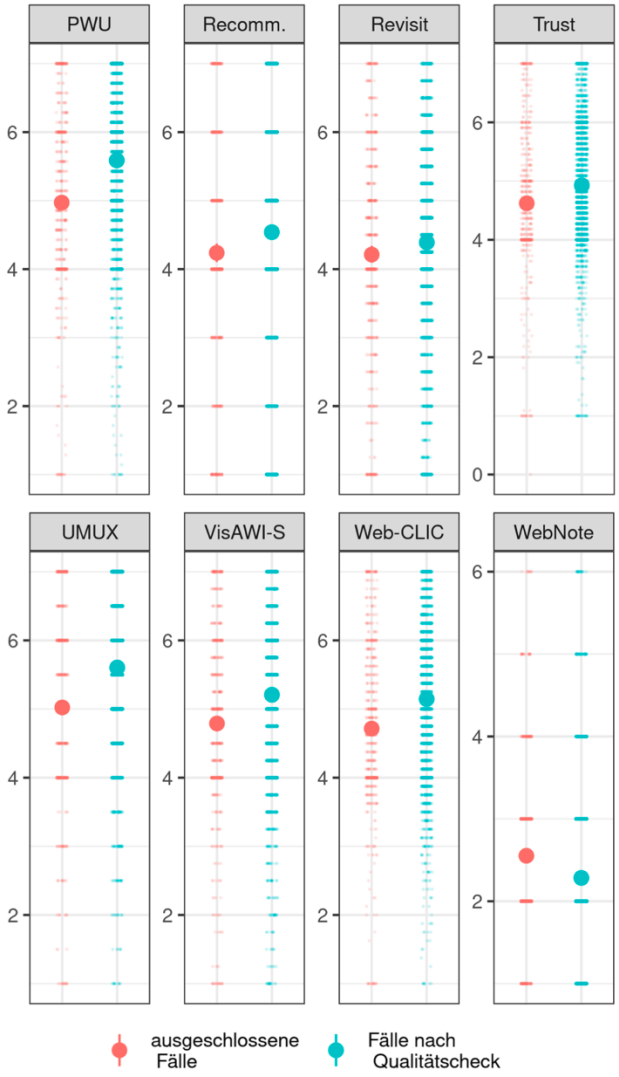

Abbildung 1: Vergleich der Mittelwerte verschiedener User Experience Indikatoren hinsichtlich der ausgeschlossenen versus der in die finale Auswertung eingeschlossenen Daten.

3 Maßnahmen zur Kontrolle der Datenqualität

Das Praxisbeispiel zeigt zum einen, welches Ausmaß Unaufmerksamkeit, Durchklicken oder bewusste Falschangaben in einer User Experience Studie auf Basis einer Panel-Befragung annehmen können. Zum andern wird ersichtlich, anhand welcher Indikatoren fehlerhafte und nicht zu verwendende Datensätze identifiziert werden können:

1. ***Freiwilliger Selbstausschluss*** am Ende der Umfrage: Manche Personen geben dort ehrlich an, dass Sie nicht ernsthaft teilgenommen oder fehlerhafte Daten erzeugt haben und schließen selbst ihre Daten aus der Auswertung aus. Die Frage nach der

Datenverwendung am Ende der Studie kann auch mit einem offenen Anmerkungsfeld kombiniert werden. Auch hier geben dann manche Befragten offen zu erkennen, wenn sie größere Verständnisprobleme mit der Studie hatten oder nicht adäquat teilgenommen haben.

2. Die ***Bearbeitungsdauer*** hat eine große Bedeutung: Durch eine Prüfung, wie viel Zeit die Teilnehmer*innen zur Beantwortung der Fragen insgesamt benötigt haben, lassen sich viele fehlerhafte Datensätze und nicht-ernsthafte Teilnahmen (= Speedster) identifizieren. An dieser Stelle lohnt sich eine realistische Schätzung, wie viel Zeit die valide Bearbeitung einer Umfrage mindestens in Anspruch nimmt. Gerade bei Usability-Testaufgaben gibt es oft Zeitvorgaben oder Erfahrungswerte wie lange diese dauern – und bei typischen Likert-Items erscheint es schon sehr merkwürdig, wenn einzelne Befragte diese kontinuierlich in unter drei Sekunden pro Frage (inkl. Lesen des Items) beantworten. Des weiteren können auch die vorliegenden Bearbeitungszeiten in einem Häufigkeitsdiagramm geplottet werden, dabei setzen sich manchmal Speedster direkt visuell in der Darstellung vom Hauptfeld der real antwortenden Befragten ab. Zudem lässt sich so bei der Analyse der Antwortdauer pro einzelner Befragungsseite feststellen, wenn bspw. die Instruktionen einer Testaufgabe nicht oder nur sehr kurz gelesen wurden.

3. ***Explizite Kontrollfragen***: In längeren Befragungen sollten zudem Fragen gestellt werden, die direkt der Qualitätskontrolle dienen. Dies können auch gezielte Fragen zum Inhalt von gezeigtem Material oder der Studie selbst sein. Meade und Craig [8] empfehlen ungefähr alle 50 bis 100 Fragen ein instruiertes Testitem (z. B. „Bitte antworten Sie bei dieser Frage zur Qualitätssicherung mit stimme zu.") einzufügen. Die Autoren raten allerdings in Summe nicht mehr als drei solcher Fragen in eine Studie einzubauen.

4. Prüfung auf ***unrealistische Angaben***: Hier kann hinsichtlich auffälliger Antwortmuster und eingeschränkter Varianz (z.B. eine Person klickt immer die gleiche Antwortoption unabhängig von Art und Richtung der Frage) sowie unmöglicher Antworten (z. B. eine Internetnutzung seit mehr Jahren als dieses existiert oder eine tägliche Onlinenutzung von > 24 Stunden) geprüft werden. Wenn vorhanden können auch die Einträge in Freitextfeldern oder in ein Anmerkungsfeld zur Umfrage selbst (oftmals am Ende der Studie gegeben) genutzt werden, um offensichtliche Fehleingaben zu identifizieren.

5. Gezielte ***technische Kontrolle***: Werden in der Online-Umfrage bestimmte Aufgaben gestellt, so kann zudem im Bedarfsfall beispielsweise mittels eines JavaScripts die Verwendung unerlaubter Hilfsmittel (siehe [4]) oder auch die Ausführung der Aufgabe (z. B. das Öffnen einer Testwebsite, vgl. [15]) kontrolliert werden.

Grundsätzlich sollten bei der Datenprüfung mehrere Kriterien kombiniert werden (siehe [3] – auch in Hinblick auf weitere mögliche Prüfkriterien). Durch solch eine Kombination von Kriterien lässt sich für die meisten Befragten sicher abschätzen, wer ernsthaft an einer Studie teilgenommen hat oder wo gewollt oder ungewollt die gemachten Angaben verfälscht sind.

4 Fazit

Das hier dargestellte Beispiel hat nicht den Anspruch auf absolute Gültigkeit – die Verteilung der verschiedenen unerwünschten Verhaltensweisen in einer User Experience Evaluation könnten in anderen Panel-Befragungen, in On-Site-Studien oder anderen Erhebungskontexten anders aussehen. Aber gerade weil es nicht möglich ist verallgemeinerbare Aussagen (wie zum Beispiel „Usability kann man nicht online testen") zu machen, ist es notwendig in jeder einzelnen Studie Maßnahmen zur Sicherung der Datenqualität zu ergreifen. Grundsätzlich müssen wir uns der Tatsache stellen, dass nicht jede Angabe in unseren Studien immer korrekt und sinnvoll auswertbar ist. Eine Prüfung der Datenqualität ist daher in User Experience Studien unerlässlich.

ACKNOWLEDGMENTS

Wir danken Simon Eisbach für die Bereitstellung des JavaScripts zur Kontrolle der Aufgabenerfüllung und David Kahre für seine Hilfe bei der Vorbereitung der Studie. Zudem sind die Autoren dankbar für die Gewährung einer Zuwendung zur Durchführung der Studie: Diese Forschung wird von der Bundeszentrale für gesundheitliche Aufklärung (BZgA) im Auftrag des Bundesministeriums für Gesundheit gefördert.

REFERENCES

[1] M. Bosnjak (2001). Classifying response behaviors in web-based surveys. Journal of Computer-Mediated Communication, 6(3), 1–14.

[2] S. Clifford, & J. Jerit (2014). Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies. Journal of Experimental Political Science, 1(2), 120–131. DOI: https://doi.org/10.1017/xps.2014.5.

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

[4] B. Diedenhofen, & J. Musch (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. Behavior Research Methods, 49(4), 1444–1459. DOI: https://doi.org/10.3758/s13428-016-0800.

[5] C. Flavián, M. Guinalíu, & R. Gurrea (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. Information & Management, 43(1), 1–14. DOI: https://doi.org/10.1016/j.im.2005.01.002.

[6] L. Küchler, G. Hertel, & M. T. Thielsch (2020). Are you willing to donate? Relationship between perceived website design, trust and donation decisions online. In: Mensch und Computer 2020 – Tagungsband (p. 223-227). ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3404983.3409993.

[7] J. R. Lewis, B. S. Utesch, & D. E. Maher (2013, April). UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 2099-2102). ACM.

[8] A. W. Meade, & S. B. Craig (2012). Identifying careless responses in survey data. Psychological Methods, 17(3), 437–455.

[9] D. H. McKnight, V. Choudhury, & C. Kacmar (2002). Developing and validating trust measures for e-commerce: An integrative typology. Information Systems Research, 13(3), 334-359.

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

[11] M. Moshagen, & M. T. Thielsch (2013). A short version of the visual aesthetics of websites inventory. Behaviour & Information Technology, 32(12), 1305-1311. DOI: doi:10.1080/0144929X.2012.694910.

[12] S. Nestler, M. T. Thielsch, E. Krasteva, & M. D. Back (2015). Will They Stay or Will They Go? Personality Predictors of Dropout in an Online Study. International Journal of Internet Science, 10 (1), 37-48.

[13] M. T. Thielsch, R. Engel, & G. Hirschfeld (2015). Expected usability is not a valid indicator of experienced usability. PeerJ Computer Science, 1-19. DOI: https://doi.org/10.7717/peerj-cs.19.

[14] M. T. Thielsch, & G. Hirschfeld (2019). Facets of website content. Human-Computer Interaction, 34(4), 279-327. DOI: https://doi.org/10.1080/07370024.2017.1421954.

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

[16] M. T. Thielsch, C. Thielsch, & G. Hirschfeld (2019). How informative is informative? Benchmarks and optimal cut points for E-Health Websites. In: F. Steinicke, & K. Wolf (Hrsg.), *Mensch und Computer 2019 – Workshopband* (p. 448-452). Gesellschaft für Informatik e.V., Bonn, Germany. DOI: https://doi.org/10.18420/muc2019-ws-642.

[17] H. Zhou, & A. Fishbach (2016). The pitfall of experimenting on the Web: How unattended selective attrition leads to surprising (yet false) research conclusions. Journal of Personality and Social Psychology, 111(4), 493–504. DOI: https://doi.org/10.1037/pspa0000056.