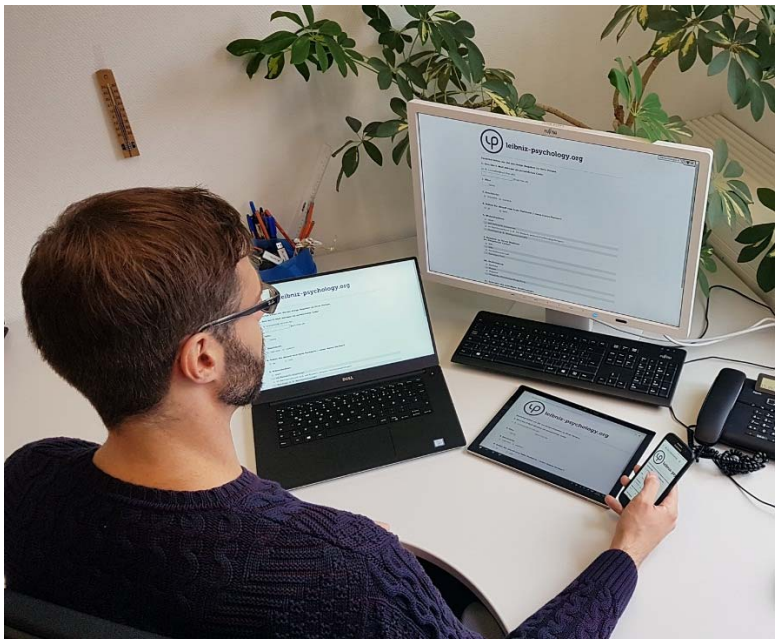
Usability-Studie zur Weiterentwicklung des ZPID-Testarchivs

Vergleich von drei neuen Webseite-Mockups mit den aktuellen Webseiten

Tom Rosman, Stefanie Mueller & Gülay Karadere



Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
Universitätsring 15
54296 Trier
www.leibniz-psychology.org

ZPID Science Information Online, 19 (1)
doi: https://doi.org/10.23668/psycharchives.2588

Usability-Studie zur Weiterentwicklung des ZPID-Testarchivs

Vergleich von drei neuen Webseite-Mockups mit den aktuellen Webseiten

Abstract

Das Elektronische Testarchiv des ZPID ist ein Open-Access-Repositorium für bisher unveröffentlichte Forschungsinstrumente aus der Psychologie. Der vorliegende Bericht beschreibt eine Nutzerstudie, die darauf abzielte, die ab 2018 durchgeführte Neuentwicklung des elektronischen Testarchivs empirisch zu begleiten. Dazu wurden drei mögliche Weiterentwicklungen der Webseiten (in Form von Mockups) hinsichtlich ihrer Nutzerfreundlichkeit und Bedienbarkeit geprüft und mit den alten Webseiten verglichen. Im Rahmen eines between-subjects-Designs mit 4 Experimentalgruppen wurden *N* = 128 Versuchspersonen aufgefordert, eine Reihe von Navigations- und Rechercheaufgaben mithilfe der Webseiten zu bearbeiten. Die jeweilige Webseite, auf der die Aufgaben zu bearbeiten war, wurde dabei experimentell und mit randomisierter Zuweisung von Personen zu Bedingungen variiert. Zur Einschätzung der Nutzerfreundlichkeit und Bedienbarkeit wurden sowohl objektive Kriterien wie die Dauer bis zur Aufgabenlösung als auch subjektive Kriterien wie die selbsteingeschätzte Nutzerfreundlichkeit herangezogen. Im Rahmen der Datenauswertung zeigte sich eine deutliche Überlegenheit der drei Mockups gegenüber den alten Seiten; zudem zeigte sich eine leichte Präferenz für die Mockup-Version 2. Diese Ergebnisse dienen als Grundlage für die Finalisierung der Neuentwicklung des elektronischen Testarchivs.

Einleitung

Das Elektronische Testarchiv ist ein Open-Access-Repositorium für bisher unveröffentlichte Forschungsinstrumente (Paper-Pencil-Verfahren) aus verschiedenen Sachgebieten und allen Bereichen der Psychologie. 1999 wurde es auf der ZIPD-Homepage freigegeben. Es ist eines der größten Testarchive in den deutschsprachigen Ländern und wird laufend aktualisiert und erweitert. Die insgesamt 194 Testinstrumente (Stand: Juni 2019) sind urheberrechtlich geschützt, stehen unter der Creative Commons Lizenz und sollen der Forschung, Lehre und Praxis dienen. Sie sind frei zum Download verfügbar. Das Elektronische Testarchiv wird kontinuierlich aktualisiert und erweitert.

Im Rahmen der strategischen Neuausrichtung des ZPID hin zu einem Universalanbieter für forschungsbasierte Infrastrukturangebote in der Psychologie wird das Elektronische Testarchiv derzeit umfassend erneuert. Vorrangig haben diese Bestrebungen zum Ziel, die Attraktivität, Funktionalität und Nutzerfreundlichkeit des Testarchivs zu steigern. Als weiteres Ziel gilt es, die Indexierung durch externe Suchmaschinen zu optimieren, was in einem weiteren Schritt erfolgen wird. Mit der neuen Webpräsenz (geplant im 3. Quartal 2019) sollen die Ziele erreicht werden.

Die Arbeitsgruppe Infrastruktur-Nutzungsszenarien begleitet diese Weiterentwicklung mit Blick auf die Usability der neuen Webseiten intensiv. Daher stellt der vorliegende Beitrag eine Usability-Studie zur empirischen Prüfung der Nutzerfreundlichkeit und Bedienbarkeit von drei Mockups der neuen Webseiten des Elektronischen Testarchivs des ZPID vor. Auf Grundlage der Studie soll eines dieser drei Mockups ausgewählt und hin zu den endgültigen neuen Webseiten des Testarchivs weiterentwickelt werden.

doi: https://doi.org/10.23668/psycharchives.2588

Methode

Webseite und Aufgaben. Im Rahmen der angesprochenen Weiterentwicklung des Elektronischen Testarchivs wurden drei Webseiten-Mockups entwickelt, die jeweils eine unterschiedliche Navigationsstruktur aufweisen. Beispielsweise war Version 2 (siehe Abb. 2) der hierarchischen Navigation des ZIS (Zusammenstellung sozialwissenschaftlicher Items und Skalen) vom GESIS (Leibniz-Institut für Sozialwissenschaften) angelehnt, während in Version 1 (siehe Abb. 1) eine klassischere, seitenbasierte Navigation implementiert wurde, die der aktuellen Testarchiv-Seiten ähneln soll. Version 3 (siehe Abb. 3) kann als Hybrid der beiden anderen Versionen angesehen werden. Version 4 (siehe Abb. 4) stellt hingegen die aktuellen (Stand: Juli 2019) Testarchiv-Seiten dar. Erstellt wurden die Mockups im Bereich I1 (Informieren und Recherchieren) des ZPID durch Dipl.-Psych. Gülay Karadere, unterstützt durch einen externen Webentwickler.

Abb. 1: Screenshot der Startseite der Mockup-Version 1

doi: https://doi.org/10.23668/psycharchives.2588

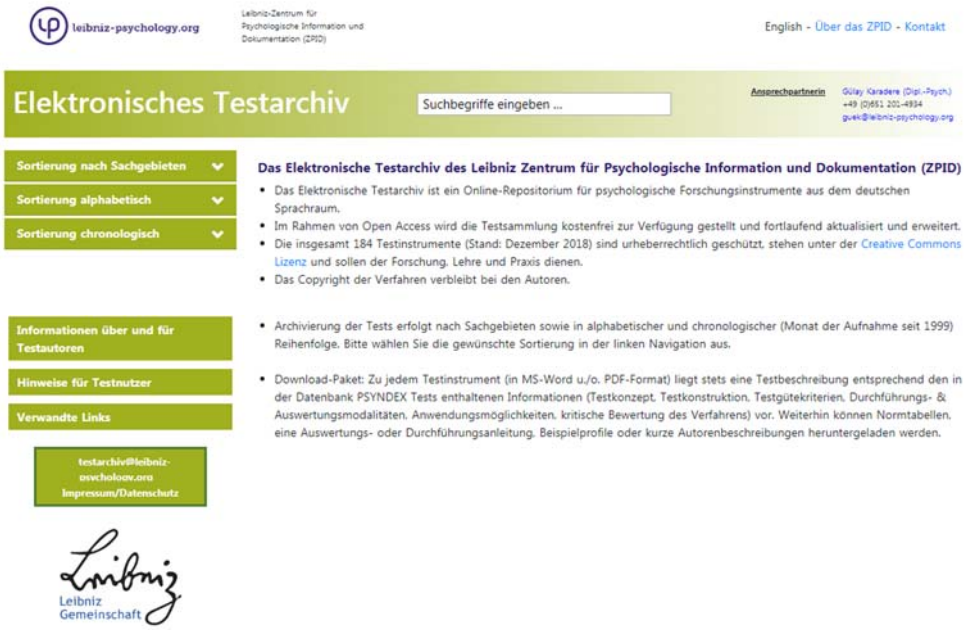


Abb. 2: Screenshot der Startseite der Mockup-Version 2

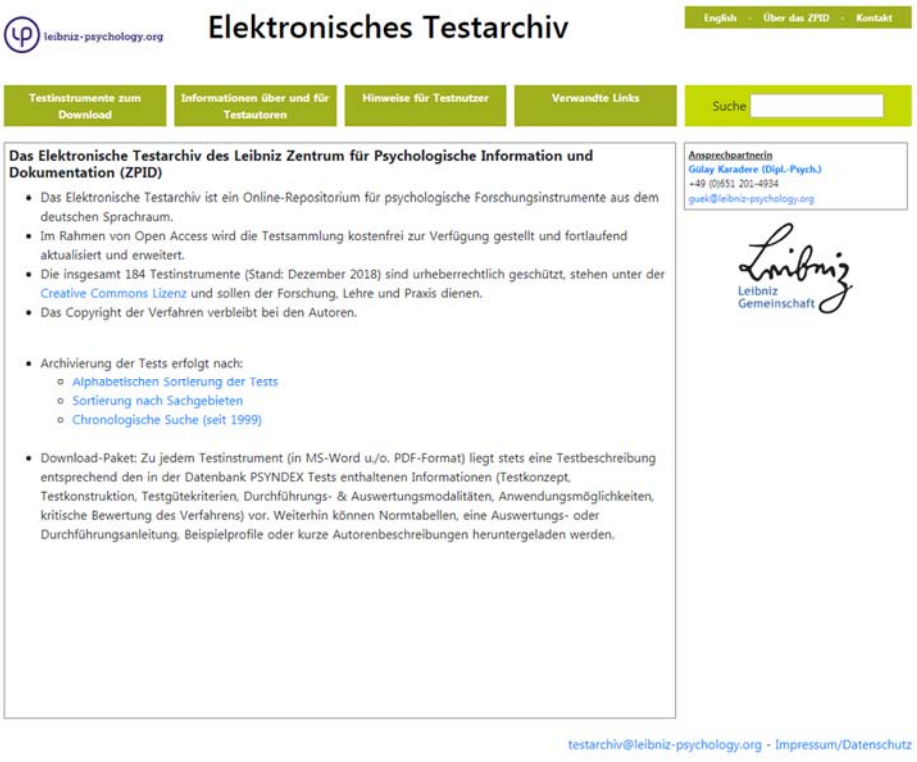


Abb. 3: Screenshot der Startseite der Mockup-Version 3

doi: https://doi.org/10.23668/psycharchives.2588

Abb. 4: Screenshot der Startseite der alten Seiten (Stand: September 2019)

Mit dem Ziel einer umfassenden Usability-Analyse wurden drei Usability-Aspekte betrachtet: (1) Die Effektivität und (2) die Effizienz der Webseiten zur Zielerreichung, sowie (3) die von potenziellen Nutzerinnen und Nutzern selbsteingeschätzte Usability. Dazu wurden insgesamt 10 Aufgaben entwickelt, die ein möglichst breites Spektrum an Aktivitäten, die auf der Webseite möglich sind, abdecken. Aufgaben 1-6 sind dabei klassische Suchaufgaben, wo die Versuchspersonen beispielsweise aufgefordert werden, den Test „Reizdarmfragebogen“ (RDF; Schäfer et al., 2018) auf den Seiten zu finden. Aufgabe 7 ist auch als Suchaufgabe zu verstehen, ist aber ökologisch valider angelegt, da die Kriterien zur Zielerreichung deutlich unschärfer sind („Suche einen Test zur Glücksforschung!“). Aufgabe 8 behandelt die Suche nach dem Rückmeldebogen, mithilfe dessen Testnutzerinnen und -nutzer den Testautorinnen und -autoren eine Rückmeldung über den Test geben können, und Aufgabe 9 behandelt die Suche nach einem Ansprechpartner oder einer Ansprechpartnerin. In Aufgabe 10 geht es schließlich um die Suche nach inhaltlichen Detailinformationen zu einem bestimmten Test - hier müssen die Versuchspersonen also einen Schritt weitergehen und die auf den Seiten verlinkte detaillierte Testbeschreibung aufrufen. Beim Design der Aufgaben wurde sichergestellt, dass alle Aufgaben auf allen Versionen der Webseite bearbeitbar sind, die Lösung also unabhängig von der jeweiligen Webseite zu finden ist.

Die Aufgaben wurden im Rahmen einer Online-Befragungssoftware (Unipark) administriert. Jede Aufgabenstellung enthielt einen Link, der auf die jeweilige Webseite führte, auf welcher die Aufgabe dann zu bearbeiten war. Nach jeder Aufgabe wurden die Versuchspersonen gebeten, die Lösung der Aufgabe in einem speziellen Feld anzugeben (z. B. sollen sie den Link zum Volltext eines bestimmten Tests angeben). Diese Lösungen wurden im Anschluss an die Studie anhand eines standardisierten Lösungsschlüssels hinsichtlich ihrer Korrektheit gerated. Zudem wurden die Versuchspersonen nach jeder Aufgabe auf einer 6-stufigen Likert-Skala um eine Einschätzung darüber gebeten, wie übersichtlich sie die Webseite während der Aufgabenbearbeitung fanden, und als wie schwierig bzw. einfach sie das Lösen der Aufgabe mithilfe der Webseite fanden. Diese Selbsteinschätzungsseite enthielt darüber hinaus ein Freitextfeld, innerhalb dessen die Versuchspersonen (freiwillig) Probleme schildern oder Verbesserungsvorschläge machen konnten.

Pilotstudie. Im Rahmen einer Pilotstudie im Bereich I2 (Studienplanungs-, Datenerhebungs- und Datenanalysedienste, Bereichsleiterin Dr. Stefanie Müller) des ZPID, wurden die Webseiten und Aufgaben einer Stichprobe von insgesamt neun Studierenden vorgelegt. Ziel dieser Studie war die Testung der Entwürfe hinsichtlich ihrer Funktionalität sowie die Identifikation von Inkonsistenzen und Unklarheiten in den Aufgabenstellungen. Zum Einsatz kamen Methoden wie Interviews, Eye-Tracking und *Lautes Denken*. Außerdem wurde die Anzahl getätigter Klicks/Schritte mit der Anzahl von Klicks/Schritten, die minimal nötig sind, um das gewünschte Ziel vom aktuellen Ausgangspunkt aus zu erreichen, verglichen (Effizienzmaß). Auf Grundlage der Daten und Rückmeldungen aus der Pilotstudie wurden diverse kleinere Optimierungen in den Aufgabenstellungen sowie in den Mockups vorgenommen.

Studiendesign und Procedere. Die Hauptstudie wurde in einem between-subjects-Design realisiert. Den Versuchspersonen wurden je nach Gruppe entweder ein Mockup der geplanten „neuen“ Testarchiv-Seiten oder aber die „alten“ Webseiten präsentiert. Die zu bearbeitenden Aufgaben, Fragen und Instrumente waren dabei für alle 4 Gruppen gleich – die Gruppen unterschieden sich lediglich hinsichtlich des ihnen zugeteilten Webseiten-Links.

Die Studie fand im PC-Pool der Psychologie der Universität Trier statt. Zu Beginn erhielten die Studierenden einen Link zu der Unipark-Umfrage. Nach einer allgemeinen Einleitung über den Zweck der Studie und einem Informed-Consent-Formular wurden einige demographische Daten abgefragt. Anschließend folgten in teilweise randomisierter Reihenfolge die 10 oben genannten Aufgaben inkl. Lösungseingabe und Selbsteinschätzungs-Abfrage. Nach Abschluss der Aufgaben wurden die Versuchspersonen schließlich noch gebeten, einige generelle Einschätzungen über die Webseite, u. A. im Rahmen einer Freitexteingabe, zu geben. Die Unipark-Umfrage war dabei selbsterklärend angelegt. Bei spezifischen Fragen konnten sich die Versuchspersonen allerdings auch bei der Versuchsleiterin melden - inhaltliche Fragen durften dabei aber nicht beantwortet werden.

Stichprobe. Eine a-priori Stichprobenumfangsplanung mittels GPower 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) ergab, dass bei einem einfaktoriellen 4-Gruppen-Design insgesamt $N = 128$ Versuchspersonen ($n = 32$ Personen pro Gruppe) nötig sind, um einen mittleren Effekt ($f = 0.30$) aufzudecken ($\alpha = .05$, $1-\beta = .80$). Entsprechend wurden im Rahmen der Studie insgesamt $N = 126$ Versuchspersonen (87.5 % weiblich; Altersdurchschnitt: $M = 23.77$; $SD = 3.65$) rekrutiert. Als Incentivierung wurde am Ende der Datenerhebung ein Versuchspersonenhonorar von 10 Euro

doi: https://doi.org/10.23668/psycharchives.2588

ausgezahlt. Alle Versuchspersonen studierten Psychologie (60.3 % B.Sc.; 34.9 % M.Sc.; 4.8 % Nebenfach); der durchschnittliche Studienfortschritt lag bei *M* = 5.65 (*SD* = 3.27) Semestern. 61.1 Prozent der Versuchspersonen gaben an, bereits einmal nach psychologischen Tests gesucht zu haben; allerdings kannten nur 27 Prozent das Testarchiv des ZPID, und nur 10.3 Prozent hatten schon einmal im Testarchiv nach Tests gesucht.

Messinstrumente. Neben aufgabenspezifischen Indikatoren (u. A. Effektivität und Effizienz der Aufgabenbearbeitung; siehe oben) wurden am Ende der Befragung drei Single-Items zu Aufbau („Die Webseite hat mir vom Aufbau her gut gefallen."), Design („Die Webseite hat ein ansprechendes Design.") und Navigierbarkeit („Die Webseite ist einfach zu navigieren.") administriert, die auf einer sechsstufigen Likert-Skala von 1 („lehne vollständig ab") bis 6 („stimme vollständig zu") zu bearbeiten waren. Zudem wurde die in diesem Kontext sehr etablierte System Usability Scale (SUS; Bangor, Kortum, & Miller, 2008; Brooke, 1996) in einer eigens vom ZPID ins Deutsch übersetzte Version (inkl. Rückübersetzung) genutzt (Beispielitem: „Ich finde, dass die verschiedenen Funktionen der Webseite gut integriert sind"; Likert-Skala von „1 = lehne vollständig ab" bis „5 = stimme vollständig zu"). Zur Auswertung wurde auf den Lösungsschlüssel von Brooke (1996) zurückgegriffen. Die Skala nimmt entsprechend einen Wertebereich zwischen 0 und 100 an. Nach Brooke (1996) gelten dabei Werte ab 50 als „ok", ab 73 als „gut" und ab 85 als „exzellent". Zur Analyse der Reliabilität wurde im Rahmen der vorliegenden Studie die interne Konsistenz (Cronbachs Alpha) herangezogen. Mit einem Wert von $\alpha = .94$ ist die Reliabilität des SUS sehr hoch.

Ergebnisse

Aus Darstellungsgründen stellen die folgenden Abschnitte nur einen Auszug aller durchgeführten Analysen dar – für eine vollständige Ergebnisdarstellung sei an dieser Stelle auf den Datensatz und die Auswertungssyntax im ZPID-Repositorium PsychArchives verwiesen.

Lösungswahrscheinlichkeit der Aufgaben (Effektivität). Zunächst wurde geprüft, inwiefern sich die Häufigkeit korrekter Lösungen je nach Webseite unterscheidet. In einem ersten Schritt wurde die Korrektheit der Aufgabenlösungen mithilfe des vorhin dargestellten Lösungsschlüssels kodiert. Eine anschließend durchgeführte einfaktorielle ANOVA (Faktor: Bedingung) ergab signifikante Gruppenunterschiede ($p < .001$) hinsichtlich des Summenwerts korrekt gelöster Aufgaben (1 Punkt pro korrekt gelöster Aufgabe, max. 10 Punkte; Deskriptivstatistiken siehe Tabelle 1). Tukey HSD Post-Hoc-Tests zeigten eine Überlegenheit aller drei neuen Seiten im Vergleich zu den alten Seiten ($p < .001$); es konnten jedoch keine signifikanten Unterschiede zwischen den drei neuen Seiten gefunden werden. Ein exakt identisches Muster zeigte sich auch für den Summenwert korrekt gelöster *Such*aufgaben (nur Aufgaben 1-6, 1 Punkt pro korrekt gelöster Aufgabe, max. 6 Punkte). Zur Testung von Gruppenunterschieden in der Lösungswahrscheinlichkeit von Aufgaben 7-10 wurden aufgrund der nicht gegebenen Intervallskalierung der abhängigen Variable (0 = falsch; 1 = korrekt) vier einzelne Kruskal-Wallis-H-Tests gerechnet. Hier wurden signifikante Gruppenunterschiede hinsichtlich der Rückmeldebogen-, Ansprechpartner- und Binnensuche-Aufgabe gefunden (mind. $p < .01$), während bezüglich der Glücksaufgabe keine signifikanten Gruppenunterschiede gefunden wurden. Da die Normalverteilungsannahme bei sämtlichen Korrektheitsindikatoren im Sinne eines Deckeneffekts verletzt ist, sind jedoch alle Ergebnisse mit Vorsicht zu interpretieren.

doi: https://doi.org/10.23668/psycharchives.2588

Tabelle 1: Deskriptive Unterschiede im Gesamtscore korrekt gelöster Aufgaben je nach Bedingung.

Version	*N*	*M*	*SD*
Version 1	31	9.19	0.91
Version 2 (ZIS)	31	9.48	0.57
Version 3	32	9.25	1.05
Alte Seiten	32	7.25	2.65
Gesamt	126	8.79	1.76

Anmerkungen. N = Stichprobengröße, *M* = Mittelwert, *SD* = Standardabweichung.

Geschwindigkeit der Aufgabenlösung (Effizienz). Anschließend wurde die Geschwindigkeit der Aufgabenlösung, ein zentraler Indikator für die Eignung der Webseite zu einer effizienten Aufgabenbearbeitung, untersucht. Dazu wurden Zeitstempel der Survey-Software Unipark analysiert. Im Rahmen dieser Analysen wurden alle Personen, die die jeweilige Aufgabe nicht lösen konnten, ausgefiltert, und zudem wurden Ausreißerwerte, die bei der Analyse von Bearbeitungszeiten schnell zu einer Verzerrung der Koeffizienten führen können, eliminiert (Kriterium: $z > 3.29$ und $z < -3.29$). Mit Ausnahme der sog. Glücksaufgabe (Aufgabe 7), für die keine signifikanten Effekte gefunden wurden, zeigten sich im Rahmen von einfaktoriellen ANOVAS kombiniert mit Tukey HSD Post-Hoc-Tests durchweg deutliche Präferenzen für die neuen Seiten im Sinne einer kürzeren Dauer bis zur korrekten Aufgabenlösung – nur einige wenige Paarvergleiche wurden nicht signifikant. Zwischen den drei neuen Webseiten zeigten sich hingegen keine signifikanten Unterschiede, wenngleich deskriptivstatistisch eine leichte Präferenz für Version 2 zu erkennen ist. Exemplarisch sei an dieser Stelle die Dauer bis zur Lösung aller Aufgaben genannt, die in Tabelle 2 dargestellt ist. Es ist allerdings darauf hinzuweisen, dass die Ergebnisse nur eingeschränkt interpretierbar sind, da in diese Darstellung nur Personen einfließen konnten, die alle Aufgaben korrekt gelöst haben, was insbesondere in Gruppe 4 mit starken Einschränkungen hinsichtlich Stichprobengröße einhergeht.

Tabelle 2: Deskriptive Daten (Mittelwert und Standardabweichung) hinsichtlich der Geschwindigkeit der Aufgabenlösung je nach Bedingung.

Version	*N*	*M*	*SD*
Version 1	13	584.15	36.93
Version 2 (ZIS)	16	570.69	19.84
Version 3	18	604.22	37.47
Alte Seiten	6	805.00	47.53
Gesamt	53	611.91	19.67

Anmerkungen. N = Stichprobengröße, *M* = Mittelwert, *SD* = Standardabweichung.

doi: https://doi.org/10.23668/psycharchives.2588

Anzahl unnötiger Navigationsschritte (Effizienz). In einem weiteren Schritt wurde geprüft, inwiefern die Versuchspersonen je nach Webseite dazu tendieren, unnötige Navigationsschritte zu machen (z. B. Navigieren hin zu Seiten, die irrelevant für die Aufgabenlösung sind). Dies kann als Indikator dafür verstanden werden, dass bestimmte Aspekte der Webseite (z. B. Links, Buttons oder Navigationsfelder) missverständlich oder ungenügend beschriftet sind. Dazu wurde auf Logdateien der Webanalyse-Software MATOMO zurückgegriffen, um das Lösungsverhalten von 86 Versuchspersonen im Rahmen der RDF-Aufgabe zu untersuchen. Alle diese Personen hatten die RDF-Aufgabe korrekt gelöst. In einem ersten Schritt wurde festgelegt, wie viele Navigationsschritte (also Aufrufe von Unterseiten) zu einer idealen Lösung der Aufgabe benötigt werden (Seite 1: 7 Schritte, Seite 2: 6 Schritte, Seite 3: 7 Schritte; alte Webseiten: 4 Schritte). Um einen über alle Bedingungen hinweg vergleichbaren Indikator zu schaffen, wurde anschließend eine nominalskalierte Variable angelegt, die indiziert, ob die jeweilige Person die Aufgabe im Rahmen der intendierten Anzahl an Navigationsschritten lösen konnte, oder ob sie zusätzliche Schritte benötigt hat. Die entsprechende gruppenspezifische Deskriptivstatistik befindet sich in Abbildung 1. Anschließend wurde ein Chi-Quadrat-Test berechnet, um zu prüfen, inwiefern sich die Häufigkeit von zusätzlichen Schritten je nach Gruppe unterscheidet. Bei diesen Analysen ergaben sich keine signifikanten Effekte. Allerdings sind auch diese Ergebnisse wiederum mit Vorsicht zu interpretieren, da die Navigation auf den alten und neuen Seiten sehr unterschiedlich ist. Beispielsweise verfügen die alten Testarchiv-Seiten über nur wenige Unterseiten, während die neuen Seiten deutlich stärker strukturiert sind. Zudem, wie aus Abbildung 1 hervorgeht, war die Stichprobengröße in den Gruppen mit unnötigen Schritten extrem gering.

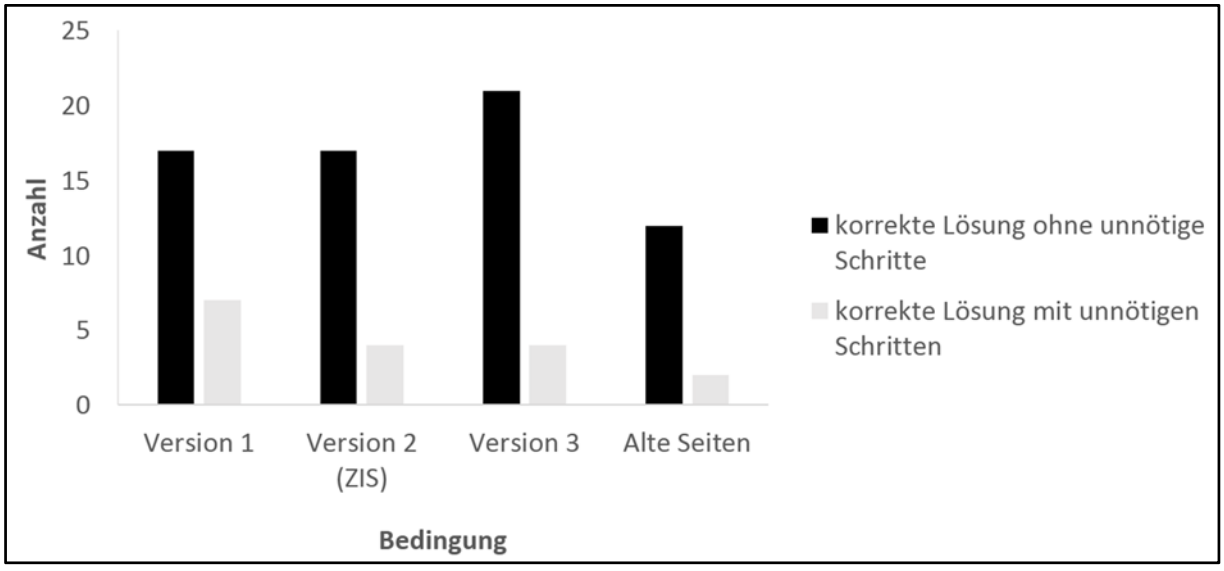


Abbildung 1: Inzidenz unnötiger Schritte je nach Bedingung.

Subjektive Einschätzungen. Abschließend wurden die subjektiven Einschätzungen der Versuchspersonen je nach Bedingung verglichen. Aus Gründen der Übersichtlichkeit wurden die Werte auf der Variable *Einfachheit der Aufgabenbearbeitung*, die für jede Aufgabe einzeln erhoben worden waren, in Form eines Mittelwerts aggregiert. Gleiches gilt für die Variable *Übersichtlichkeit der Webseite bei der Aufgabenbearbeitung*. Zusätzlich zu diesen beiden Maßen wurden die SUS-Skala

doi: https://doi.org/10.23668/psycharchives.2588

sowie die Single-Items zu Aufbau, Design und Navigierbarkeit der Webseite analysiert. Eine deskriptivstatistische Darstellung dieser Ergebnisse befindet sich in Tabelle 3. Im Rahmen von einfaktoriellen ANOVAS zeigte sich eine starke Präferenz der Versuchspersonen für die neuen Seiten, die auf allen Maßen auch auf den jeweiligen Tukey HSD Post-Hoc-Tests statistisch hoch signifikant war ($p < .001$). Wie bei den vorher berichteten Tests waren die Unterschiede zwischen den einzelnen Webseiten deutlich geringer: Lediglich auf den Items zu Aufbau, Design und Navigierbarkeit zeigte sich eine leichte Präferenz für Version 2. So wurde Version 2 im Vergleich zu Version 1 und 3 als marginal signifikant ($p < .10$) besser aufgebaut eingeschätzt. Bezüglich Design schnitt Version 2 zudem signifikant besser als Version 1 ($p < .05$) ab, während es zwischen Version 3 und Version 1 keine signifikanten Unterschiede im Design gab. Schließlich schnitt Version 2 auch hinsichtlich Navigierbarkeit besser ab als Version 1, während keine entsprechenden Unterschiede zwischen Version 3 und Version 1 gefunden wurden.

Zusammenhänge zwischen den Indikatoren. Um zu prüfen, inwiefern die einzelnen hier berichteten Indikatoren zusammenhängen, wurden Pearson-Korrelationen zwischen aggregierten Maßen für (1) die Effektivität (Korrektheit der Aufgabenbearbeitung) und (2) die Effizienz der Aufgabenbearbeitung (Dauer der Aufgabenbearbeitung) sowie (3) die subjektiv eingeschätzte Güte der Webseite berechnet[1]. Dabei zeigte sich eine sehr hohe Korrelation zwischen dem subjektiven Maß und der Effektivität der Aufgabenbearbeitung ($r = .67$; $p < .001$), während nur geringe, aber dennoch erwartungskonforme Zusammenhänge zwischen Effizienz (Geschwindigkeit der Aufgabenbearbeitung) und dem subjektiven Maß ($r = -.23$; $p < .05$) sowie zwischen Effizienz und Effektivität gefunden wurden ($r = -.22$; $p < .05$). Allerdings sind auch diese Ergebnisse aufgrund der nicht gegebenen Normalverteilung der Effektivitäts-Variable als vorläufig anzusehen.

[1] Für eine genaue Darstellung der Variablenaggregation sei an dieser Stelle auf die Auswertungssyntax verwiesen.

doi: https://doi.org/10.23668/psycharchives.2588

Tabelle 3: Deskriptive Unterschiede in den subjektiven Einschätzungen der Webseiten je nach Bedingung.

Abhängige Variable	Version	*N*	*M*	*SD*
SUS	Version 1	31	80.65	16.60
	Version 2 (ZIS)	31	89.19	10.67
	Version 3	32	85.63	11.05
	Alte Seiten	32	59.61	21.61
Übersichtlichkeit bei der Aufgabenbearbeitung	Version 1	31	4.63	0.85
	Version 2 (ZIS)	31	5.01	0.79
	Version 3	32	4.86	0.71
	Alte Seiten	32	3.67	0.86
Einfachheit der Aufgabenbearbeitung	Version 1	31	4.83	0.77
	Version 2 (ZIS)	31	5.15	0.38
	Version 3	32	5.05	0.59
	Alte Seiten	32	4.04	0.83
Aufbau	Version 1	31	4.97	0.95
	Version 2 (ZIS)	31	5.55	0.57
	Version 3	32	4.97	0.86
	Alte Seiten	32	3.81	1.35
Design	Version 1	30	3.70	1.47
	Version 2 (ZIS)	31	4.77	1.02
	Version 3	32	4.47	1.24
	Alte Seiten	32	3.38	1.50
Navigierbarkeit	Version 1	31	4.71	1.27
	Version 2 (ZIS)	31	5.48	0.68
	Version 3	32	5.28	0.85
	Alte Seiten	32	3.69	1.23

Anmerkungen. N = Stichprobengröße, *M* = Mittelwert, *SD* = Standardabweichung. SUS = System Usability Scale (Bangor, Kortum, & Miller, 2008; Brooke, 1996):Wertebereich: 0-100; 3 Cut-off-Werte: 50-72 = „ok“, 73-84 = „gut“, 85-100 = „exzellent“.

doi: https://doi.org/10.23668/psycharchives.2588

Fazit

Aus den vorliegenden Ergebnissen lassen sich zwei Schlüsse ziehen. Zunächst einmal legen sie auf beeindruckende Weise die Überlegenheit der drei Mockups im Vergleich zu den alten Seiten dar. Liegt der SUS-Score der alten Seiten noch bei moderaten 59 Punkten, wird die Usability der neuen Seiten mit einem mittleren Score von 85 („exzellent“) deutlich besser bewertet. Mit Blick auf diese ermutigenden Ergebnisse ist zudem zu bedenken, dass die fest eingeplante weitere Verbesserung des final ausgewählten Mockups hinsichtlich Layout (z. B. Anpassung des Farbschemas) und Funktionalität (z. B. Implementation der Suchfunktion) noch weitere Usability-Verbesserungen nach sich ziehen wird.

Trotzdem gibt es einen Wermutstropfen. Die Usability der drei neuen Mockups ist scheinbar so gut, dass die vorliegenden statistischen Analysen, womöglich aufgrund von Deckeneffekten auf den jeweiligen abhängigen Variablen, nur wenig Unterschiede zwischen den drei Mockup-Versionen aufzeigen konnten. Zwar ist der SUS-Wert von Version 2 mit 89.2 Punkten nochmal höher als derjenige von Version 1 (80.6 Punkte) und Version 3 (85.6 Punkte), allerdings sind diese Unterschiede in den jeweiligen Post-Hoc-Tests statistisch nicht signifikant. Lediglich auf den drei Single-Items zur Selbsteinschätzung der Usability zeigt sich eine statistisch einigermaßen abgesicherte Präferenz für Version 2. Auch bei Betrachtung der Deskriptivstatistiken der übrigen Indikatoren fällt auf, dass Version 2 auf fast allen Indikatoren leicht überlegen ist – jedoch dicht gefolgt von Version 3. Zusammengenommen lässt sich somit in den vorliegenden Daten eine leichte Präferenz für Version 2 erkennen, aber auch eine Wahl von Version 3 für die finalen Webseiten ist aus der Sicht der Autor/-innen dieses Beitrags wissenschaftlich vertretbar.

Endgültige Klarheit in dieser Frage könnte lediglich eine Wiederholung der Datenerhebung mit einem noch größeren Sample schaffen, allerdings ist hervorzuheben, dass eine Stichprobengröße von *N* = 126 Personen auch jetzt schon deutlich über dem Durchschnitt gewöhnlicher Usability-Studien liegt. Zudem kann die geplante qualitative Auswertung der diversen Freitextantworten noch weitere Klarheit bringen. Insgesamt liefert die vorliegende Studie damit einen wertvollen Beitrag zur Weiterentwicklung des Elektronischen Testarchivs und damit zum Infrastrukturauftrag des ZPID.

Literatur

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24* (6), 574–594. https://doi.org/10.1080/10447310802205776

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland, I. L. (Ed.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41* (4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Schäfer, S. K., Weidner, K. J., Becker, N., Stokes, C. S., Lammert, F., & Köllner, V. (2017). RDF. Reizdarmfragebogen [Fragebogen]. In Leibniz-Zentrum für Psychologische Information und

Dokumentation (ZPID) (Hrsg.), Elektronisches Testarchiv (PSYNDEX Tests-Nr. 9007494). Trier: ZPID. https://doi.org/10.23668/psycharchives.826