# Stichprobenumfang und Fragebogenlänge in Webevaluationen

**Gerrit Hirschfeld**
gerrit.hirschfeld@fh-bielefeld.de
Fachhochschule Bielefeld
Bielefeld, NRW

**Meinald T. Thielsch**
thielsch@uni-muenster.de
Westfälische Wilhelms-Universität Münster
Münster, NRW

## ABSTRACT

Bei der Planung von Studien zur Evaluation von Websites stellen sich in der Praxis häufig zwei zentrale Fragen: (1) Welches Instrument soll zur Evaluation verwendet werden? (2) Wie viele NutzerInnen müssen befragt werden? Die erste Frage ist dabei durch das Evaluationsziel der jeweiligen Studie bestimmt, zur zweiten lassen sich jedoch allgemeine Aussagen treffen. Im vorliegenden Beitrag wird der Zusammenhang zwischen der erforderlichen Anzahl an Befragten und der Reliabilität der verwendeten Fragebögen aus psychometrischer Perspektive dargestellt und diese beiden Größen miteinander in Beziehung gesetzt. Basierend auf Befragungen zum Website-Content wird untersucht, wie viele NutzerInnen mit unterschiedlich langen Fragebögen befragt werden müssen, um eine gewünschte Präzision zu erzielen. Aus den Ergebnissen werden konkrete Tipps abgeleitet, die StudienplanerInnen berücksichtigen sollten, wenn sie Evaluationsstudien planen.

## CCS CONCEPTS

• **General and reference → Cross-computing tools and techniques**; **Reliability**; **Estimation**.

## KEYWORDS

Studienplanung, Stichprobenumfang, Fragebogenlänge, Webevaluationen

**ACM Reference Format:**
Gerrit Hirschfeld and Meinald T. Thielsch. 2019. Stichprobenumfang und Fragebogenlänge in Webevaluationen. In *Mensch und Computer 2019 (MuC '19), September 8–11, 2019, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340764.3344454


*MuC '19, September 8–11, 2019, Hamburg, Germany*

ACM ISBN 978-1-4503-7198-8/19/09…$15.00
https://doi.org/10.1145/3340764.3344454

## 1 EINLEITUNG

Die Website-Evaluation hat sich aus der Evaluation von Softwareprodukten, einem Spezialthema der Mensch-Maschine Interaktion [7], entwickelt. Bei der Evaluation von Webseiten stellen sich in der Praxis zwei Fragen:

(1) Welches Instrument soll ich verwenden, um die NutzerInnen zu befragen?
(2) Wie viele Probanden müssen befragt werden?

Die Antwort auf die erste Frage ergibt sich aus den jeweiligen Evaluationszielen und dem Fokus einer Website- Evaluation. Die Festlegung einer klaren Fragestellung und konkreter Evaluationsziele ist entscheidend für den Projekterfolg. Daraus ergibt sich auch, welche Evaluationskriterien und damit welche spezifischen Befragungsinstrumente herangezogen werden sollten. Im nächsten Schritt ist konkret zu überlegen, welche Methoden notwendig sind, um die Evaluationsfrage zu beantworten. Obwohl anhand vielfältiger neuer Technologien von Logfile-Analysen bis hin zu Deep Learning immer neue Methoden zur Bewertung von Webseiten herangezogen werden können, ist die Evaluation durch potentielle und/oder tatsächliche NutzerInnen anhand von Fragebögen immer noch eine häufig verwendete und zentrale Informationsquelle zur Qualitätseinschätzung einer Website. Aus praktischer Sicht ist vor allem die Länge des Fragebogens, d.h. die Menge an zu bearbeitenden Items kritisch, weil sie beeinflusst, wie viele NutzerInnen rekrutiert werden können. So ist es viel leichter, NutzerInnen zur Teilnahme an sehr kurzen Erhebungen mit wenige Items zu gewinnen als umfangreiche Fragebögen beantworten zu lassen [1]. Ein Nachteil kurzer Instrumente ist allerdings deren geringere Reliabilität, die dazu führt, dass das interessierende Merkmal ungenauer erfasst wird [6].

Im vorliegenden Beitrag soll aus psychometrischer Sicht genau dieser Zusammenhang hergestellt werden, um für eine geplante Evaluationsstudie die optimale Kombination aus Stichprobenumfang und Fragebogenlänge zu finden. Um diese Zusammenhänge darzustellen, haben wir exemplarisch die Domäne der Bewertung von Webseiten hinsichtlich ihres Inhalts anhand des Web-CLIC-Fragebogens verwendet. Der Inhalt ist das zentrale Element einer Website – und wird von NutzerInnen als das wichtigste Kriterium für die

Beurteilung genannt [9]. Der Web-CLIC ist ein Fragebogen, mit dem NutzerInnen den Inhalt einer Website bewerten können [8]. Insgesamt 12 Fragen decken dabei die vier Bereiche Verständlichkeit, Gefallen, Informationsgehalt und Glaubwürdigkeit ab (die Abkürzung Web-CLIC steht für Website-Clarity, Likeability, Informativeness, Credibility). Zudem kann ein Gesamtwert berechnet werden, dieser spiegelt das subjektive Erleben des Website-Inhalts insgesamt wider. Beispielhafte Items des Web-CLICs sind:

- Die Inhalte sind anschaulich aufbereitet. (Skala: Verständlichkeit)
- Ich lese diese Website gerne. (Skala: Gefallen)
- Die Website ist informativ. (Skala: Informationsgehalt)
- Ich kann den Informationen auf der Website vertrauen. (Skala: Glaubwürdigkeit)

Die Konstruktion und Validierung des Web-CLICs basieren auf sechs Studien mit insgesamt n = 3106 Befragten und m = 60 getesteten Websites [8]. Der Fragebogen erweist sich als reliabel: Die interne Konsistenz (Cronbachs $\alpha$) ist $\geq .80$ für die Skalen und $\geq .90$ für den Gesamtwert. Die Zeitstabilität (Retest-Reliabilität) über einen Zeitraum von 2 Wochen liegt für die Skalen im Bereich von $.69 \leq r \leq .81$, bzw. bei $r = .84$ für den Gesamtwert. Hinsichtlich der Validität findet sich eine umfassende empirische Evidenz. Geprüft wurden faktorielle, konvergente, divergente, diskriminative, konkurrente, experimentelle und prädiktive Validität.

## 2 METHODE

### Minimale Stichprobenumfänge

Oftmals hört man in der Praxis die von Jacob Nielsen geprägte Daumenregel fünf Testpersonen seien ausreichend [5]. Allerdings können bei nur fünf Testpersonen aber substanzielle Probleme übersehen werden. Faulkner empfahl daher die Zahl auf mindestens 10, für sichere Datengrundlagen auf 20 anzuheben [3]. Derartige Daumenregeln sind hilfreich für explorativ-qualitative Usability-Tests, aber hinsichtlich einer fragebogenbasierten Website-Evaluation lässt sich der notwendige Stichprobenumfang statistisch genau bestimmen:

Um einen minimal erforderliche Stichprobenumfang abzuschätzen, muss man bestimmte Annahmen machen und Zielgrößen definieren. Im Falle von Evaluationsstudien von Webseiten ist es das Ziel, möglichst genau ein bestimmtes Merkmal (etwa die Bewertung des Inhalts) zu schätzen. Diese Genauigkeit kann man am einfachsten anhand der Breite des Konfidenzintervalls beschreiben. Konfidenzintervalle geben einen Bereich um den empirisch gefundenen Mittelwert an, innerhalb dessen in 95% der Fälle der wahre Wert liegt. Konfidenzintervalle werden mit zunehmenden Stichprobenumfang kürzer und somit genauer, liegen also näher um den wahren Wert. Umgekehrt bedeutet dies, dass man mehr Probanden braucht, je genauer man das Merkmal schätzen möchte. Neben der gewünschten Genauigkeit ist auch die Variabilität des Merkmals relevant. Je stärker dieses Merkmal schwankt, desto mehr Probanden müssen befragt werden. Unter der Annahme der Normalverteilung besteht der folgende Zusammenhang zwischen der Standardabweichung eines Merkmals, der gewünschten Präzision (gemessen als Breite des Konfidenzintervalls der Mittelwerte) und dem erforderlichen Stichprobenumfang [4]:

$$n = \frac{z_{1-\frac{\alpha}{2}} * S^2}{e^2}$$

Wobei $\alpha$ dem gewünschten Konfidenzniveau (z.B. 95%), $e$ der halben Länge des gewünschten Konfidenzintervalls und $S$ der Standardabweichung der Messwerte entspricht.

### Tatsächliche und beobachtete Standardabweichungen

Die Standardabweichung der Messwerte wiederum setzt sich zusammen aus der Reliabilität und der tatsächlichen Standardabweichung [10]. Aus der Definition der Reliabilität als Anteil der wahren Varianz an der Varianz der gemessenen Werte ergibt sich folgende Formeln für die Varianz der Messwerte ($\sigma_x^2$):

$$\sigma_x^2 = \frac{\sigma_X^2}{p_{xx}}$$

Wobei $\sigma_X^2$ der Varianz des Merkmals und $p_{xx}$ der Reliabilität des Messinstruments entspricht. Je höher die Reliabilität des Messinstruments desto geringer ist die Varianz der gemessenen Werte.

### Abhängigkeit der Reliabilität von der Itemanzahl

Zuletzt nutzt man den Zusammenhang zwischen der Itemanzahl und der Reliabilität des Messinstruments. So kann man anhand der Spearman-Brown-Vorhersageformel [6] schätzen, wie sich die Reliabilität eines Fragebogens verändert, wenn man die Länge variiert.

$$Rel_{neu} = \frac{k * Rel_{alt}}{1 + (k-1)Rel_{alt}}$$

Wobei $Rel_{alt}$ die Ausgangsreliabilität und $k$ den Kürzungsfaktor (k = Anzahl Items der neuen Version / Anzahl der Items der Ausgangsversion) beschreibt.

## 3 ERGEBNISSE

### Zusammenhang zwischen beobachteter Standardabweichung, Präzision und Stichprobenumfang

Abbildung 1 zeigt die benötigten Probanden, um Skalen mit Standardabweichungen von 0,9; 1; 1,3 und 2 unterschiedlich

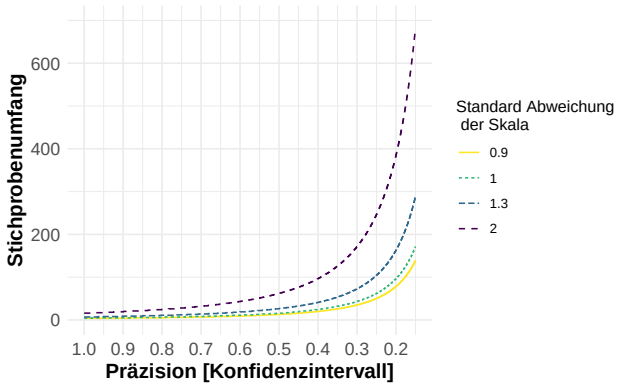


**Figure 1: Erforderliche minimale Stichprobenumfänge in Abhängigkeit zur gewünschten Präzision**

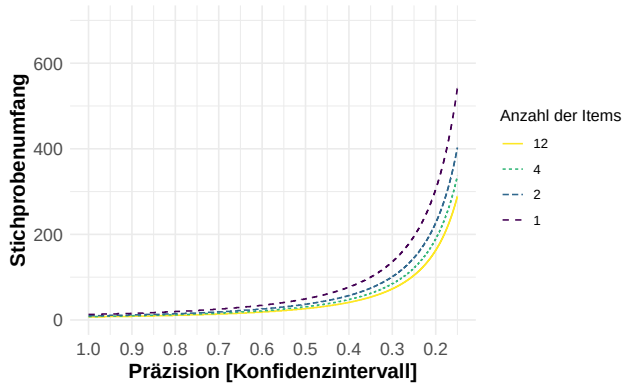


**Figure 2: Minimal erforderliche Stichprobenumfänge für unterschiedliche Anzahlen von Items und unterschiedlich hohe Präzision.**

genau zu schätzen. Eine Präzision von 1 bedeutet dabei, dass der wahre Wert nicht weiter als einen Skalenpunkt vom Mittelwert der Ratings entfernt liegt. Ist der Mittelwert der Ratings also bei 4, kann man nur schließen, dass der wahre Wert zwischen 3 und 5 liegt. Das ist für die meisten praktischen Anwendungen wahrscheinlich wenig informativ. Oft ist es daher sinnvoll, eine Präzision von weniger als 0,15 Punkten anzustreben, um den wahren Wert deutlich genauer schätzen zu können – im obigen Beispiel läge der wahre Wert dann zwischen 3,85 und 4,15. Um dies beispielsweise für den Web-CLIC, der eine Standardabweichung von 1,3 aufweist, zu erreichen, benötigt man 290 Probanden.

### Korrigierte Standardabweichung der Ratings

Im Falle des Web-CLIC liegt die Reliabilität des Gesamtscores bei 0,92 (Cronbach's $\alpha$) und die beobachtete Standardabweichung bei 1,3. Daraus folgt, dass die korrigierte Standardabweichung des Web-CLIC bei 1,25 Skalenpunkten auf einer 7-stufigen Skala liegt. Zwar unterscheiden sich aufgrund der hohen Reliabilität die beiden Schätzungen für die Standardabweichung kaum. Allerdings kann nun mit Hilfe der korrigierten Standardabweichung berechnet werden, wie hoch die Standardabweichung der Messwerte für unterschiedlich reliable Skalen wird. Verwendet man zum Beispiel eine Skala, die nur eine Reliabilität von 0,5 aufweist, um das Merkmal zu erfassen, würde nach obiger Formel die Standardabweichung der Messwerte 2,5 betragen.

### Zusammenhang zwischen Fragebogenlänge, gewünschter Präzision und erforderlichem Stichprobenumfang

Im letzten Schritt verwenden wir nun die Spearman-Brown Formel, um die Reliabilität von unterschiedlich langen Skalen zur Erfassung des Webseiteninhalts zu schätzen. Wir vergleichen dafür auf Basis des Web-CLICs vier verschiedene hypothetische Inhaltsfragebögen mit einer Länge von 1, 2, 4 und 12 Items. Legt man wieder wie im Web-CLIC eine Reliabilität von 0,92 für den 12-Item Fragebogen zugrunde, ergeben sich geschätzte Reliabilitäten für die o.g. Kurzversionen von 0,50; 0,66; 0,76, und 0,92. Diese kann man nun verwenden, um die Standardabweichung der Ratings zu berechnen, aus denen man dann die Anzahl der benötigten Probanden ablesen kann.

Abbildung 2 stellt die erforderlichen Stichprobenumfänge graphisch dar. Wie man sieht, unterscheiden sich die erforderlichen Stichprobenumfänge vor allem dann, wenn die gewünschte Präzision relativ hoch ist. Um ein Konfidenzintervall mit einer Breite von einem Skalenpunkt (= Präzision von 0,5) zu erreichen, sind je nach Anzahl der Items entweder 27 Probanden bei 12 Items oder 49 Probanden bei nur einem Item erforderlich.

## 4 DISKUSSION

Das Ziel des Beitrags war es, die Zusammenhänge zwischen Stichprobenumfang und Fragebogenlänge aus psychometrischer Sicht darzustellen. Die Ergebnisse zeigen dabei, dass der erforderliche Stichprobenumfang in erster Linie von der gewünschten Präzision - oder umgekehrt: der tolerierten Unsicherheit - abhängt. Möchte man Inhaltsratings auf einen Skalenpunkt genau schätzen, genügen bei einer Skala mit 12 Items bereits 26 Probanden, während man im selben Fall 104 Probanden benötigt, um Ratings auf einen halben Skalenpunkt genau zu schätzen. Der zweite Faktor ist die Länge des Fragebogens bzw. die damit einhergehende Reliabilität. Dieser ist umso wichtiger, je höher die gewünschte Präzision ist. Gleichzeitig nimmt der Zugewinn durch längere Fragebögen auch immer mehr ab. Das heißt, die Verbesserung von einem Item auf drei Items ist größer als die Verbesserung von 12 auf 14.

Aus praktischer Sicht ergeben sich die folgenden Empfehlungen für die Erfassung von Ratings in Befragungen:

- PraktikerInnen sollten im Rahmen der Studienplanung festlegen, wie genau die Ratings das Merkmal erfasst werden soll. Als Daumenregel erscheint es sinnvoll, eine Präzision von einem halben Skalenwert auf siebenstufigen Skalen anzustreben.
- Bei Einzelitem-Messungen ist die benötigte Anzahl der Befragten extrem erhöht, im Vergleich sind bereits Kurzskalen mit nur drei Items bereits deutlich besser.
- Werden Merkmale mit ähnlichen wahren Standardabweichung wie Inhaltsratings (SD = 1,25) untersucht, sollten bei Skalen mit einer Reliabilität von 0,7/0,8/0,9 mindestens 196/151/119 Probanden befragt werden, um eine Präzision von einem halben Skalenpunkt zu erzielen.

Einige Einschränkungen sind hierbei jedoch relevant. Erstens basieren die berechneten Stichprobenumfänge auf den psychometrischen Merkmalen von Inhaltsratings von Webseiten mit dem Web-CLIC. Für andere Fragebögen, die eine höhere oder geringere Reliabilität aufweisen, und höhere gewünschte Präzisionen ergeben sich andere Werte. Möchte man beispielsweise mit der Systems Usability Scale, die eine Reliabilität von .85 aufweist, ein Konfidenzintervall mit einer Breite von .15 erreichen, sind 1477 Probanden notwendig. Dieselbe Präzision kann bei einer Reliabilität von 0,92 bereits mit 1261 Probanden erreicht werden. Eine zweite Einschränkung ist, dass zur Berechnung der Ergebnisse auf psychometrische Zusammenhänge aus der Klassischen Testtheorie (KTT) zurück gegriffen wurde. Die Ergebnisse gelten also nur insofern, als dass die Annahmen der KTT gelten. Weitere Studien sollten Simulationsstudien verwenden, um die untersuchten Zusammenhänge unter realistischeren Bedingungen zu betrachten. Drittens möchten wir darauf hinweisen, dass die Präzision nicht das einzig mögliche Zielkriterium für Evaluationsstudien ist. Gelegentlich ist das Ziel einer Evaluationsstudie, unterschiedliche Websites im Rahmen eines AB-Tests direkt miteinander zu vergleichen. Hier kann es sein, dass bereits 50 Personen ausreichen, um mittel-große Unterschiede zwischen den Websites zu detektieren. Die exakte Größe kann – und sollte auch - in diesem Fall mit gängigen Open Source Programmen zur Stichprobenplanung, wie z.B. G*power [2], berechnet werden. Allerdings sind solche Ansätze nicht anwendbar, wenn keine alternative Webseite für einen Vergleich vorliegt. Man muss also für eine korrekte Bestimmung des Stichprobenumfangs auch das jeweilige Evaluationsziel berücksichtigen.

## 5 REFERENZEN

[1] Fan, W. and Yan, Z. 2010. Factors affecting response rates of the web survey: A systematic review. *Computers in human behavior*. 26, 2 (2010), 132–139.

[2] Faul, F. et al. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*. 39, (2007), 175–191.

[3] Faulkner, L. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*. 35, 3 (2003), 379–383.

[4] Kauermann, G. and Küchenhoff, H. 2011. *Stichproben*. Springer.

[5] Nielsen, J. 1993. *Usability engineering*. AP Professional.

[6] Nunnally, J.C. et al. 1967. *Psychometric theory*. McGraw-Hill.

[7] Salaschek, M. et al. 2007. Benutzbarkeit von Software: Vor- und Nachteile verschiedener Methoden und Verfahren. *Zeitschrift für Evaluation*. 6, 2 (2007), 247–276.

[8] Thielsch, M.T. and Hirschfeld, G. 2019. Facets of website content. *Human Computer Interaction*. 34, (2019), 279–327. DOI:10.1080/07370024.2017.1421954.

[9] Thielsch, M.T. et al. 2014. User evaluation of websites: From first impression to recommendation. *Interacting with Computers*. 26, 1 (2014), 89–112.

[10] Trafimow, D. 2014. Estimating true standard deviations. *Frontiers in Psychology*. 5, 235 (2014).