Fragebogen und ordinale Urteile

Theo Held

SAP SE

theo.held@sap.com

Zusammenfassung

Bekannterweise liefern die üblichen in Fragebogen verwendeten Ratingskalen mit n Stufen lediglich ordinal skalierte Zahlenwerte. Die Relevanz der Skalenniveaus im Sinne von Stevens (1946) ist ein vieldiskutierter Streitpunkt bei dem sich die „wahre Lehre“ der axiomatischen Messtheorie und die eher pragmatische Auffassung auf der Anwenderseite gegenüberstehen. Grundsätzlich stellt sich die Frage, welche statistischen Verfahren für die Ergebnisse solcher Ratings sinnvoll angewandt werden können und welche Konsequenzen es möglicherweise hat, Kennwerte zu verwenden für die es aus messtheoretischer Sicht keine Grundlage gibt. Außerdem soll diskutiert werden, welche alternativen oder ergänzenden Verfahren verfügbar sind, die der ordinalen Natur von Ratings Rechnung tragen.

1 Einleitung

Mehrstufige Ratings sind ein Bestandteil sehr vieler Fragebögen im Bereich von Usability und User Experience. Gemeinhin haben diese Fragebögen den Anspruch, theoretische Konstrukte wie User Experience oder Usability mit ihren vielfältigen Dimensionen zu „messen“. Die zentrale Frage ist, was diese Messung in theoretischer Hinsicht eigentlich bedeutet. Ein schöner Überblick zu dieser Fragestellung findet sich z.B. in Lewis (2014).

Allgemein bedeutet Messung im psychologischen Bereich eine Zuordnung von numerischen Werten (numerisches Relativ) zu empirischen Objekten (empirisches Relativ) mithilfe einer Abbildungsvorschrift (z.B., dass einem als intensiver empfundenen Stimulus auch eine höhere Zahl zugeordnet werden muss). Stevens (1946) postuliert, dass resultierende Zahlen im Bereich der psychologischen Messung auf unterschiedlichen Skalenniveaus einzuordnen sind. Die wichtigsten Niveaus sind: Nominale Skalierung („Etikettierung“ von empirischen Objekten), Ordinale Skalierung (Zahlen repräsentieren eine Ordnung der empirischen Objekte), Intervallskalierung (Differenzen zwischen den Zahlen repräsentieren Differenzen zwischen den empirischen Objekten) und Verhältnisskalierung (numerische Verhältnisse zwischen den Messwerten repräsentieren Verhältnisse zwischen den empirischen Objekten).

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2018 in
R. Dachselt, G. Weber (Hrsg.):
Mensch und Computer 2018 – Workshopband, 02.–05. September 2018, Dresden.
 https://doi.org/10.18420/muc2018-ws16-0492

Die numerischen Werte, die den Ratings in Fragebogen zugeordnet werden, sind entsprechend dieser Differenzierung ordinal skaliert, das bedeutet, dass z.B. bei bipolaren Fragebogenitems (z.B. „stimme nicht zu“ mit dem numerischen Wert 1 bis „stimme voll zu“ mit dem Wert 7) davon ausgegangen wird, dass ein höherer numerischer Wert einer höheren Zustimmung entspricht, dass aber keine Aussagen bzgl. der einzelnen Intervalle zwischen den Bewertungsstufen gemacht werden können und dass insbesondere Aussagen wie „eine Zustimmung mit dem numerischen Wert 4 ist doppelt so stark wie eine Zustimmung mit dem Wert 2“ nicht zulässig sind.

2 Maß der zentralen Tendenz für Praktiker

Die Frage ist nun, was man prinzipiell mit diesen Werten statistisch/rechnerisch machen kann. Streng genommen ist die Verwendung des arithmetischen Mittels nicht zulässig, da grundlegende mathematische Operationen auf diesen ordinal skalierten Zahlen zwar möglich, aber nicht sinnvoll sind (bei ordinal skalierten Zahlen sind alle ordnungserhaltenden Transformationen zulässig, somit ist es völlig egal, ob eine Reihe empirischer Objekte durch die Zahlen 1, 2, 3 oder 5, 200, 3800 repräsentiert wird). Somit verbieten sich grundsätzlich auch statistische Vergleiche der Mittelwerte, z.B. mit t-Tests.

In statistischer Hinsicht liegt es deshalb nahe, als Maß der zentralen Tendenz nicht das arithmetische Mittel, sondern den Median zu verwenden und zum Vergleich der Verteilungen von Fragebogenergebnissen ein non-parametrisches Verfahren wie den Mann-Whitney U Test (Mann & Whitney, 1947), der auch keine Normalverteilung der Daten voraussetzt.

Jedoch ist die Berechnung des arithmetischen Mittels durchaus gängige Praxis und es ist auch üblich per t-Tests, Varianzanalysen, etc. statistische Mittelwertvergleiche durchzuführen. Die Frage ist, welche potenziellen Auswirkungen das haben kann, bzw. welches Risiko diese „lockere“ Sichtweise bzgl. des zugrundeliegenden Skalenniveaus in sich birgt. Auch die umgekehrte Frage stellt sich: könnte es sein, dass unter bestimmten Bedingungen die Verwendung der Mittelwerte und parametrischer Verfahren sogar Vorteile bietet? Lewis (1993) stellt dazu für zwei Studien die Verwendung parametrischer (t-Test) und non-parametrischer Verfahren (Mann-Whitney U Test) zur Analyse von Fragebogendaten gegenüber. Verglichen werden die Resultate des „Post-Study System Usability Questionnaire“ (PSSUQ; Lewis, 1992) mit 7-stufigen Ratings und des „After-Scenario Questionnaires“ (ASQ; Lewis, 1991) mit 5-stufigen Ratings. Von besonderem Interesse war dabei, wie stark die ermittelte statistische Signifikanz der Unterschiede zwischen Mittelwerten bzw. Medianen mit den Mittelwerten bzw. Medianen an sich zusammenhängt. Ausgangspunkt dieser Untersuchung war die konkrete Beobachtung, dass es durchaus vorkommen kann, dass die Mediane für zwei Bedingungen praktisch gleich sein können wobei jedoch das non-parametrische Verfahren durchaus einen signifikanten Unterschied zwischen den Bedingungen zeigt. Solche Fälle treten z.B. dann auf, wenn die Verteilungen der untersuchten Datensätze eine stark gegenläufige Schiefe aufweisen (Lewis, 1993).

Für den Praktiker, der Resultate solcher fragebogenbasierten Untersuchungen zur Unterstützung von Management-Entscheidungen berichten muss, ergibt sich in einem solchen Fall ein Problem. Die Maße der zentralen Tendenz werden üblicherweise als die wesentlichen Indikatoren für Unterschiede kommuniziert. Wie die Arbeit von Lewis (1993) zeigt, gibt es Fälle, in denen der statistisch signifikante Unterschied zwischen Fragebogenergebnissen besser durch die arithmetischen Mittel als durch die Mediane repräsentiert wird.

Ein weiteres Ergebnis der Studie von Lewis (1993) war, dass die Mittelwerte die aus einem 7-stufigen Rating (PSSUQ) resultieren die ermittelte Signifikanz der Mittelwertunterschiede besser repräsentieren als die aus 5-stufigen Ratings (ASQ) resultierenden Mittelwerte. Dies entspricht auch den Ergebnissen von Nunnally (1978), nach denen die die Reliabilität mehrstufiger Skalen im Bereich von 2 bis 7 Stufen stark ansteigt.

Bei der Verwendung mehrstufiger Ratings müssen auch stets potenzielle Effekte berücksichtigt werden, die dadurch entstehen, dass die Diskriminationsmöglichkeit im Bereich der Maximalstufe sehr eingeschränkt ist. Solche Deckeneffekte führen naturgemäß dazu, dass Mittelwerte und Standardabweichungen wegen der Deckelung des Maximalwerts ebenfalls gedeckelt sind bzw. zu niedrig geschätzt werden. Neben einer Kombination von Ratings mit Verfahren, die diese Einschränkung nicht aufweisen (siehe Abschnitt 3.3) existieren auch Möglichkeiten, die Verteilungsparameter a-posteriori zu korrigieren. Alliger et al. (1988) stellen dazu einen Ansatz zur Korrektur von Deckeneffekten vor und untersuchen die Verwendbarkeit mit Realdaten und in einer Simulationsstudie.

Grundsätzlich ist es für den Praktiker wichtig, sich der messtheoretisch begründeten Problematik bewusst zu sein und die Stärke der resultierenden Aussagen nicht in unzulässiger Weise überzustrapazieren. Aussagen bzgl. des numerischen Verhältnisses der Mittelwerte sind definitiv nicht zulässig (Wert A ist doppelt so gut wie Wert B, etc.).

Man sollte sich zudem neben Mittelwert oder Median immer die Verteilung der Daten genauer ansehen. Für den Praktiker ergeben sich hier wichtige weitere Anhaltspunkte für mögliche Unterschiede.

3 Alternative und kombinierte Methoden

Speziell bei kritischen Entscheidungen auf Basis von Fragebogenergebnissen sollte man zusätzliche/alternative methodische Ansätze ins Auge fassen.

Als Alternativen oder Ergänzungen bieten sich methodische Vorgehensweisen an, die a-priori auf der ordinalen Ebene angesiedelt sind. Dazu zählen z.B. präferenzorientierte Verfahren wie Rangreihungen (Rankings) oder Paarvergleiche. Diese Verfahren haben wiederum spezifische ökonomische Nachteile bzgl. ihrer Durchführbarkeit – andererseits weisen sie auch eine Reihe von Vorteilen auf: sie können potenziell zu verhältnisskalierten Resultaten führen, einfache Rangordnungsverfahren sind für die Teilnehmer einfach durchzuführen und sie bieten eine gute Gelegenheit, bereits während der Durchführung auch inhaltliche Gründe für die einzelnen Präferenzen einzusammeln.

3.1 Paarvergleiche

Die Methode der vollständigen Paarvergleiche folgt einem einfachen Prinzip. Im Laufe einer Beurteilungssequenz werden alle Paare aus unterschiedlichen zu beurteilende Objekten dargeboten. Der Beurteiler hat sich bei jedem Paar für eines der beiden Objekte in Abhängigkeit von einem vorgegebenen Kriterium zu entscheiden. Kriterien können simple physikalisch fundierte Kriterien, wie z.B. Größe, Lautstärke oder Gewicht sein, oder aber abstraktere Eigenschaften wie Attraktivität, Wichtigkeit, oder nur eine nicht weiter begründete Präferenz. Es werden also rein ordinale Urteile gefordert.

Paarvergleiche bieten den Vorteil, dass die Konsistenz der Urteile einfach überprüft werden kann. Wenn eine Person Alternative A Alternative B vorzieht (z.B. „schöner", „attraktiver", „größer") und B gegenüber Alternative C präferiert, dann ist zu erwarten, dass auch A gegenüber C präferiert wird. Interessant ist hier, zu untersuchen, ob bei unterschiedlichen Vergleichspaaren unterschiedliche Eigenschaftsdimensionen als Grundlage der Entscheidung herangezogen werden. Für eine Beschreibung der Methode und deren Einordnung im Kontext anderer Skalierungsmethoden sei z.B. auf Borg & Staufenbiel (2007) verwiesen.

Wichtig für den Praktiker ist, dass Paarvergleiche eine relativ aufwändige Methode sind, da es für n Objekte n(n-1)/2 Paare gibt. Bei 5 Objekten sind somit mindestens 10 Vergleiche erforderlich, bei 10 Objekten sind es bereits 45. Andererseits ist die kognitive Belastung für den Beurteiler relativ gering und es bietet sich die Möglichkeit, auch Begründungen für die Entscheidungen abzufragen.

Wie auch im Falle der im nächsten Abschnitt beschriebenen Rangreihungen können natürlich Rating-basierte Fragebogenitems nicht ohne Weiteres durch ein Paarvergleichs-Verfahren ersetzt werden. Meist muss ein erweitertes Verfahren zur Anwendung kommen (siehe Abschnitt 3.3).

3.2 Ranking

Ebenso wie Paarvergleiche erfordern Rankings in ihrer Grundform rein ordinale Urteile. Die Beurteiler haben hier die Aufgabe, Objekte hinsichtlich eines gegebenen Kriteriums in eine Rangreihe zu bringen (z.B. eine Sortierung entsprechend der steigenden subjektiven Attraktivität von Nutzungsschnittstellen). Die ordinale Skalierung der Objekte ergibt sich aus den Rangzahlen innerhalb der Anordnung. Im Gegensatz zu einem Rating der Attraktivität per Objekt müssen sich hier die Beurteiler intensiv mit dem Vergleich zwischen den Objekten auseinandersetzen. Ein Vorteil dieser Vorgehensweise ist die Möglichkeit einer besseren Differenzierung zwischen Objekten, andererseits ist eine solche Anordnung durch das Verfahren (zumindest in seiner starken Form) vorgeschrieben, d.h. die Beurteiler werden ähnlich wie bei Paarvergleichen gezwungen, sich in Zweifelsfällen für ein Objekt zu entscheiden. Es sei hier nur darauf hingewiesen, dass es auch Ranking-Verfahren gibt, die unvollständige oder schwache Sortierungen mit Gruppen von Objekten mit gleicher Rangzahl vorsehen.

3.3 Ratings in Kombination mit „ordinalen“ Methoden

Ein gemeinsames Problem von Paarvergleich und Ranking ist, dass eine differenzierte Reihung der Objekte erreicht wird, jedoch keine Aussagen darüber möglich ist, wie stark die subjektive Ausprägung des Ordnungskriteriums bei einzelnen Objekten ist. Es kann z.B. sein, dass Beurteiler eine gegebene Menge von Objekten problemlos bezüglich ihrer Attraktivität anordnen können, sie aber auch das am höchsten eingestufte Objekt noch durchaus unattraktiv finden – der Rest ist einfach noch weniger attraktiv.

Eine Möglichkeit, die Vorteile von Ratings und von Rangordnungsverfahren oder Paarvergleichen zu nutzen, ist eine Kombination der Verfahren. McCarthy & Shrum (2000) stellen ein Verfahren vor, in dem zunächst ein Ranking der Objekte und dann ein Rating durchgeführt wird. Der Vorteil des Verfahrens liegt darin, dass durch das Ranking die häufig schwache Differenzierung der Rating-Ergebnisse verbessert wird.

Böckenholt (2004) macht einen Vorschlag, wie man dem Problem, dass ordinale Verfahren keinen Skalennullpunkt liefern, durch eine Kombination von Paarvergleichen und absoluten Einschätzungen wie Ratings begegnen kann.

4 Fazit

Dass die Ergebnisse von Ratings ordinal skaliert sind, aber üblicherweise wie mindestens intervallskalierte Daten behandelt werden, führt für den Nutzer von rating-basierten Fragebögen zwangsläufig zu einigen Konsequenzen:

- Die Interpretation der arithmetischen Mittel von Ratings sollte sich nicht auf die Distanz zwischen den Mittelwerten beziehen und Aussagen zu numerischen Verhältnissen der Mittelwerte sind auf jeden Fall zu vermeiden. Der Median der Ratings sollte zusätzlich zum arithmetischen Mittel berichtet werden, wobei es durchaus sein kann, dass ein in einem non-parametrischen Verfahren ermittelter signifikanter Unterschied für Berichtszwecke besser durch den Mittelwert repräsentiert wird.
- Man sollte sich neben den Maßen der zentralen Tendenz immer auch die Verteilung von Rating-Resultaten ansehen.
- Wenn möglich, sollten – zumindest für wichtige Entscheidungen – Ratings nicht als einziges Verfahren verwendet werden. Es bieten sich die oben erwähnten kombinierten Verfahren an, es sollte aber auch erwogen werden, alternative Verfahren unabhängig voneinander zu verwenden und die Resultate auf gemeinsame Tendenzen zu untersuchen.

Literaturverzeichnis

Alliger, G. M., Hanges, P. J., & Alexander, R. A. (1988). A method for correcting parameter estimates in samples subject to a ceiling. *Psychological Bulletin, 103*(3), 424-430.

Borg, I. & Staufenbiel, T. (2007). *Lehrbuch Theorien und Methoden der Skalierung* (4. Auflage). Hans Huber, Hogrefe AG: Bern.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9,* 453-465.

Lewis, J.R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin, 23,* 78-81.

Lewis, J.R. (1992). Psychometric evaluation of a post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting* (S. 1259-1263). Santa Monica, CA: Human Factors Society.

Lewis, J.R. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction, 5,* 383-392.

Lewis, J.R. (2014). Usability: Lessons learned … and yet to be learned. *International Journal of Human-Computer Interaction, 30*(9), 663-684.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics. 18*(1), 50–60.

McCarty, J.A. & Shrum, J.J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly, 64,* 271-298.

Nunnally, J.C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103,* 677-680.

Autor

Held, Theo

Dr. Theo Held studierte Psychologie an der Universität Regensburg. Nach der Promotion (Universität Heidelberg) war er in Forschung und Lehre an den Universitäten Heidelberg, Graz und Halle/Saale tätig. Seine Forschungsinteressen liegen in den Bereichen Wahrnehmung und Wissensrepräsentation, sowie der Evaluation von Softwareprodukten. Seit 2001 gehört er dem User Experience Team der SAP an. Bis Ende 2010 war er für eine Reihe zentraler Designkonzepte der SAP CRM Lösung verantwortlich. Seit 2011 ist er als User Experience Researcher tätig.