Messung wahrgenommener Benutzbarkeit: Vergleich von verbalen und piktoralen Ein-Item- und Multi-Item-Fragebogen

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Elisa Katharina Gräve

aus Düsseldorf

Düsseldorf, Oktober 2024

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Axel Buchner

2. Prof. Dr. Raoul Bell

Tag der mündlichen Prüfung: 08.01.2025

Inhaltsverzeichnis

Zusammenfassung

Für die Entwicklung und Bewertung interaktiver Computersysteme ist es neben der Erfassung objektiver Maße der Benutzbarkeit auch wichtig, die wahrgenommene Benutzbarkeit zu evaluieren. Sie wird beschrieben durch die subjektive Leichtigkeit, mit der Menschen mit einem System interagieren. Um kosteneffizient umfangreiche Daten über die wahrgenommene Benutzbarkeit zu erheben, werden oft standardisierte Fragebogen in unterschiedlicher Länge und Komplexität verwendet. In der einschlägigen Literatur zur Messung psychologischer Konstrukte mit Fragebogen werden Fragebogen mit mehr Items gegenüber Fragebogen mit weniger Items mehrheitlich bevorzugt. Jedoch haben sich in vielen Bereichen Ein-Item-Fragebogen als valide Messinstrumente erwiesen. In der vorliegenden Dissertation wurde geprüft, ob das psychologische Konstrukt der wahrgenommenen Benutzbarkeit besser durch Fragebogen mit einer größeren oder mit einer kleineren Anzahl an Items gemessen werden kann. Dazu wurde in den Experimenten 1a und 1b experimentell getestet, wie gut verbale Fragebogen – der 35-Item ISONORM 9241/10 (nur Experiment 1a), die 10-Item System Usability Scale, die 4-Item Usability Metric for User Experience und die Ein-Item Adjective Rating Scale – den Unterschied in der wahrgenommenen Benutzbarkeit zwischen zwei simulierten webbasierten Mobiltelefonvertragssystemen reflektieren. Die beiden Systeme wurden für die Experimente so entworfen, dass sie sich in ihrer Benutzbarkeit offensichtlich unterscheiden. Die Teilnehmenden arbeiteten mit einem der beiden Systeme und bewerteten danach ihre Erfahrungen mit dem jeweiligen System anhand aller Fragebogen. Es zeigte sich, dass die Ein-Item Adjective Rating Scale den Unterschied in der wahrgenommenen Benutzbarkeit zweier Systeme mindestens genauso gut oder sogar signifikant besser als die längeren Fragebogen reflektierte. Dieses Ergebnis konnte bezüglich des Vergleichs der Adjective Rating Scale und der System Usability Scale in den Experimenten 2a und 2b repliziert werden. Außerdem zeigte sich ein vergleichbares Befundmuster für eine piktorale Alternative, die Pictorial Single-Item Usability Scale, die den Unterschied in der wahrgenommenen Benutzbarkeit zwischen zwei Systemen mindestens genauso gut reflektierte wie die längere 10-Item Pictorial System Usability Scale. Es lässt sich im Rahmen dieser Dissertation also schlussfolgern, dass sowohl die verbale Ein-Item Adjective Rating Scale als auch die Ein-Item Pictorial Single-Item Usability Scale als valide und effiziente Maße für die Messung von Gesamtunterschieden in der wahrgenommenen Benutzbarkeit empfohlen werden können.

Abstract

For the development and evaluation of interactive computer systems, it is important to assess perceived usability in addition to recording objective measures of usability. Perceived usability is described as the subjective ease with which people interact with a system. In order to cost-efficiently collect extensive data on perceived usability, standardized questionnaires of varying lengths and complexity are often used. In the relevant literature on measuring psychological constructs with questionnaires, questionnaires with more items are generally favored over questionnaires with fewer items. However, single-item questionnaires have proven to be valid measurement instruments in many areas. In the present dissertation it was examined whether the psychological construct of perceived usability is better measured by questionnaires with a larger or with a smaller number of items. To this end, Experiments 1a and 1b tested how well verbal questionnaires – the 35-item ISONORM 9241/10 (Experiment 1a only), the 10-item System Usability Scale, the 4-item Usability Metric for User Experience and the single-item Adjective Rating Scale – reflect the difference in perceived usability between two simulated web-based mobile phone contract systems. The two systems were designed for the experiments in such a way that they clearly differed in their usability. The participants worked with one of the two systems and then rated their experience with the respective system using all the questionnaires. It was found that the single-item Adjective Rating Scale reflected the difference in the perceived usability of two systems at least as well or significantly better than the longer questionnaires. This result was replicated with regard to the comparison of the Adjective Rating Scale and the System Usability Scale in Experiments 2a and 2b. In addition, a comparable pattern of findings was found for a pictorial alternative, the Pictorial Single-Item Usability Scale, which reflected the difference in perceived usability between two systems at least as well as the longer 10-item Pictorial System Usability Scale. Accordingly, it can be concluded within the scope of this dissertation that both the verbal single-item Adjective Rating Scale and the Pictorial Single-Item Usability Scale can be recommended as valid and efficient measures for measuring overall differences in perceived usability.

Einleitung

In den frühen 1980er-Jahren wurde die *Ingenieurpsychologie (engl. Human Factors Psychology)* verstärkt in die Entwicklung und Bewertung kommerzieller Computersysteme miteinbezogen, da erkannt worden ist, dass es nicht ausreicht, nur objektive Maße wie Fehlerraten oder die Zeit, die für die Durchführung einer Aufgabe benötigt wird, zu erfassen. Stattdessen ist es auch wichtig, die Wahrnehmung und die Reaktionen der Menschen, die sich aus der (erwarteten) Nutzung eines Systems ergeben, zu evaluieren (ISO-9241-210, 2010). Die sogenannte *Benutzererfahrung* umfasst die gesamte Erfahrung, die Benutzende vor, während und nach der Interaktion mit einem System machen. Dabei werden auch emotionale Reaktionen, Erwartungen sowie die allgemeine Zufriedenheit mit dem System einbezogen (ISO-9241-210, 2010). Benutzererfahrungen können sowohl die Kaufabsicht (Fedele et al., 2017) als auch die Wahrscheinlichkeit, dass das genutzte System weiterempfohlen wird (Brooke, 2013), beeinflussen. Beide Faktoren stellen potenzielle Erfolgsfaktoren für ein System dar. Ein wichtiger Teil dieser Benutzererfahrung bildet das psychologische Konstrukt der *Benutzbarkeit*. Dieses wird definiert durch das Ausmaß, in dem ein Produkt von bestimmten Nutzenden verwendet werden kann, um bestimmte Ziele mit Effektivität, Effizienz und Zufriedenheit in einem bestimmten Nutzungskontext zu erreichen (ISO-9241/11, 1998).

Die Benutzbarkeit fokussiert sich auf die funktionale Handhabung eines Systems. Sie wird von Fachleuten aus der Praxis gemessen, um festzustellen, ob das entworfene System den Bedürfnissen der Benutzenden entspricht und für einen Markteintritt bereit ist. Von Anwendungsforschenden wird sie hingegen gemessen, um sicherzustellen, dass die neuen Schnittstellendesigns, an denen sie arbeiten, die natürlichen Grenzen der menschlichen Benutzenden, wie begrenzte kognitive Kapazitäten, motorische Fähigkeiten und sensorische Wahrnehmungen, nicht überschreiten oder übermäßig herausfordern (Kortum & Oswald, 2018). Ein wichtiger Aspekt der Benutzbarkeit ist das psychologische Konstrukt der *wahrgenommenen Benutzbarkeit* (Lewis, 2018), die die subjektive Komponente der Benutzbarkeit darstellt. Die wahrgenommene Benutzbarkeit beschreibt die subjektive Leichtigkeit, mit der Menschen mit einem System interagieren. Sie ist entscheidend für die Entwicklung und Optimierung interaktiver Systeme wie

Webseiten und Softwareanwendungen (Baumgartner et al., 2019b). Der pragmatischste Weg, die wahrgenommene Benutzbarkeit zu untersuchen, besteht darin, die Benutzenden einfach nach ihrer Meinung zu fragen (Nielsen, 1994). Um kosteneffizient umfangreiche Daten über die wahrgenommene Benutzbarkeit eines Systems vor oder nach seiner Markteinführung zu erheben, werden häufig sprachbasierte, also verbale standardisierte Fragebogen verwendet. Ein standardisierter Fragebogen ist ein Fragebogen, der für eine wiederholte Verwendung konzipiert ist und typischerweise eine bestimmte Reihe von Fragen in einer bestimmten Reihenfolge und in einem bestimmten Format enthält (Lewis et al., 2015).

Die ersten standardisierten verbalen Fragebogen zur Messung wahrgenommener Benutzbarkeit wurden in den späten 1980er-Jahren veröffentlicht und werden seither in der Forschung und von Fachleuten aus der Praxis eingesetzt (Lewis, 2018). Anfang der 1990er-Jahre wurde dann der 35 Items umfassende verbale *ISONORM 9241/10* (Prümper, 1993) veröffentlicht, der die direkte Operationalisierung der entsprechenden ISO-Norm darstellt und somit eine solide theoretische Grundlage aufweist (ISO-9241/10, 1995). Der ISONORM 9241/10 gilt als ein reliables und valides Instrument zur Messung der wahrgenommen Benutzbarkeit (Prümper, 1997, 1999). Zeitlich kurz nach dem ISONORM 9241/10 wurde die 10-Item *System Usability Scale* (Brooke, 1996, 2013) veröffentlicht, die mittlerweile aufgrund ihrer Popularität oft als Industriestandard angesehen wird (Brooke, 2013). Die System Usability Scale gilt als ein valides und äußerst robustes Instrument, da sie vielseitig eingesetzt werden kann, zur Bewertung der Qualität eines breiten Spektrums von Systemen und Technologietypen geeignet ist und robust gegenüber verschiedener Kontexten ist (Bangor et al., 2008). Da die Items der System Usability Scale abwechselnd positiv und negativ formuliert sind, was unter anderem zur Fehlinterpretation von negativ formulierten Items führen kann, entwickelten Sauro und Lewis (2011) eine rein positiv formulierte Alternativversion. Es konnte gezeigt werden, dass die positiv formulierte System Usability Scale und die ursprüngliche System Usability Scale im Wesentlichen die gleichen Ergebnisse liefern und sie dementsprechend in der Praxis austauschbar verwendet werden können (Kortum et al., 2021). Vor allem für nicht überwachte Online-Studien wird die positiv formulierte System Usability Scale empfohlen (Sauro & Lewis, 2011). Für Situationen mit zeitlichen Beschränkungen wurde später die kurze 4-Item *Usability Metric for*

User Experience entwickelt (Finstad, 2010). Dieser verbale Fragebogen basiert auf der ISO-9241/11 (1998)-Definition von Benutzbarkeit und die Forschung berichtet über wünschenswerte psychometrische Eigenschaften (Berkman & Karahoca, 2016; Finstad, 2010; Lewis et al., 2013, 2015). Der zeiteffizienteste Fragebogen ist ein Fragebogen mit einem Item, wie die *Adjective Rating Scale*, die Bangor et al. (2009) der System Usability Scale als elftes Item hinzugefügt haben. Die Werte der Adjective Rating Scale korrelierten dabei hoch mit den Summenwerten der System Usability Scale (r = .822). Angesichts dieser hohen Korrelation könnte die Adjective Rating Scale eine zeiteffiziente Alternative zu Fragebogen mit mehreren Items sein.

Bei der Verwendung verbaler Fragebogen kann es aufgrund von möglichen Sprachbarrieren der Befragten zu Fehlinterpretationen von Items kommen (Bradley & Lang, 1994; Kunin, 1955). Aus diesem Grund wurden piktorale Fragebogen für ein breites Spektrum von Anwendungsbereichen entwickelt, da sie aufgrund der bildlichen Darstellungen weitgehend sprachfrei und dementsprechend als leicht verständlich und intuitiv zu beantworten gelten – selbst für Kinder oder Personen, die die Sprache, in der der Fragebogen verfasst ist, nur begrenzt beherrschen (Bradley & Lang, 1994). Für die Bewertung der wahrgenommenen Benutzbarkeit entwickelten Baumgartner et al. (2019a) die *Pictorial System Usability Scale* als eine piktorale Alternative zur verbalen System Usability Scale. Eine erste Validierungsstudie lieferte vielversprechende Ergebnisse (Baumgartner et al., 2019a), da die Summenwerte der Pictorial System Usability Scale und der verbalen System Usability Scale hoch korrelierten (r = .865). Jedoch sollten aufgrund der relativ kleinen Stichprobe (N = 60) weitere Studien durchgeführt werden, um die psychometrischen Eigenschaften des Fragebogens weitergehend zu untersuchen. Für die Pictorial System Usability Scale wurde ebenfalls eine Ein-Item-Alternative zur Messung der wahrgenommenen Benutzbarkeit entwickelt, die sogenannte *Pictorial Single-Item Usability Scale* (Baumgartner et al., 2019b). Eine erste Validierungsstudie zeigte insgesamt gute Ergebnisse (Baumgartner et al., 2019b): Die Werte der Pictorial Single-Item Usability Scale korrelierten hoch mit den Summenwerten der verbalen System Usability Scale (r = .696). Aufgrund der relativ kleinen Stichprobe (N = 38) ist es jedoch auch hier notwendig, die psychometrischen Eigenschaften des Fragebogens mithilfe weiterer Studien zu prüfen.

Die vorgestellten verbalen und piktoralen Fragebogen zur Messung der wahrgenommenen Benutzbarkeit variieren deutlich in ihrer Komplexität und Länge und enthalten zwischen 35 Items und einem Item. In der einschlägigen Literatur zur Messung psychologischer Konstrukte mit Fragebogen werden Fragebogen mit mehr Items gegenüber Fragebogen mit weniger Items mehrheitlich bevorzugt, da Messungen, die auf vielen Items beruhen, oft als reliabler angesehen werden als Messungen, die auf wenigen Items oder sogar nur auf einem Item beruhen. Dies basiert auf der Annahme, dass die Verwendung vieler Items den Messfehler reduziert und somit eine präzisere Messung der psychologischen Konstrukte ermöglicht (z.B. Churchill, 1979; Nunnally, 1978). In der Praxis werden aufgrund ihrer Effizienz jedoch zunehmend kurze Fragebogen bevorzugt, da sie schnell ausgefüllt werden können und einfach zu administrieren sind (Pomeroy et al., 2001). Das Ausfüllen langer Fragebogen der wahrgenommenen Benutzbarkeit kann für Teilnehmende nach einem Tag mit vielen Benutzbarkeitstestungen ermüdend sein, da ihnen unter anderem ein hohes Maß an Motivation und Anstrengung abverlangt wird (Wanous et al., 1997). Lange Fragebogen können die Befragten sogar überfordern (Wanous et al., 1997), was zu einer Abnahme der Datenqualität und -quantität führt (Dillman et al., 1993). Ein-Item-Fragebogen haben dementsprechend pragmatisch einige Vorteile und sind daher bei gleicher Validität im Vergleich zu einer Multi-Item-Alternative zu bevorzugen. Tatsächlich werden viele verbale und piktorale Ein-Item-Maße in verschiedenen Bereichen als valide Messinstrumente angesehen, da wiederholt gezeigt wurde, dass ihre psychometrischen Eigenschaften denen ihrer Multi-Item-Pendants gleichwertig sind, wie unter anderem im Bereich der allgemeinen Arbeitszufriedenheit (Dolbier et al., 2005; Nagy, 2002; Oshagbemi, 1999; Wanous et al., 1997) oder der Werbeforschung (Ang & Eisend, 2018; Bergkvist & Rossiter, 2007, 2009).

Um zu eruieren, ob das psychologische Konstrukt der wahrgenommenen Benutzbarkeit besser durch standardisierte Fragebogen mit einer größeren oder mit einer kleineren Anzahl Items gemessen werden kann, wurden im Rahmen der vorliegenden Dissertation die oben vorgestellten Fragebogen experimentell in vier Experimenten verglichen. Genauer gesagt wurde überprüft, wie gut diese Fragebogen den Unterschied in der wahrgenommenen Benutzbarkeit zwischen einem System mit guter Benutzbarkeit und einem System mit schlechter Benutzbarkeit reflektieren. In der ersten Testreihe (Experiment 1a und 1b) wurden

die folgenden verbalen Fragebogen experimentell verglichen: der ISONORM 9241/10, die System Usability Scale, die Usability Metric for User Experience und die Adjective Rating Scale. In der zweiten Testreihe (Experiment 2a und 2b) wurden zwei verbale Fragebogen – die System Usability Scale und die Adjective Rating Scale – und zwei piktorale Alternativen – die Pictorial System Usability Scale und die Pictorial Single-Item Usability Scale – experimentell verglichen.

Vergleich von verbalen Fragebogen zur Messung der wahrgenommenen Benutzbarkeit

Das Ziel der ersten Testreihe war es, experimentell zu testen, ob verbale standardisierte Fragebogen mit einer größeren oder mit einer kleineren Anzahl Items für die Messung des psychologischen Konstrukts der wahrgenommenen Benutzbarkeit besser geeignet sind. Dafür wurde getestet, welcher verbale Fragebogen den Unterschied in der wahrgenommenen Benutzbarkeit zwischen einem System mit guter Benutzbarkeit und einem System mit schlechter Benutzbarkeit am besten reflektiert. Um sicherzustellen, dass die Teilnehmenden keine Vorerfahrungen mit einem der beiden Systeme hatten, da andernfalls die Bewertung der wahrgenommenen Benutzbarkeit beeinflusst werden könnte (Berkman & Karahoca, 2016; Borsci et al., 2015), wurden im Rahmen der Dissertation zwei neue simulierte webbasierte Mobiltelefonvertragssysteme (Web-Anwendungen zu einer simulierten Mobilfunkvertragsauswahl) entworfen, die sich in ihrer Benutzbarkeit deutlich unterschieden. Das System mit einer guten Benutzbarkeit (fortan gut gestaltetes System) wurde so entworfen, dass es einem Standard von Dialogprinzipien der Mensch-System-Interaktion entsprach (ISO-9241-110, 2008), wohingegen das System mit einer schlechten Benutzbarkeit (fortan schlecht gestaltetes System) viele dieser Prinzipien verletzte (siehe einen Auszug aus den Web-Anwendungen in Einzelarbeit 1, Abbildung 1).

Experiment 1a

In Experiment 1a wurden die vier verbalen Fragebogen, der 35-Item ISONORM 9241/10 (Prümper, 1993), die 10-Item positiv formulierte System Usability Scale (Sauro & Lewis, 2011), die 4-Item Usability Metric for User Experience (Finstad, 2010) und die Ein-Item Adjective Rating Scale (Bangor et al., 2009) experimentell

verglichen. Die Teilnehmenden arbeiteten mit einem der beiden webbasierten Mobiltelefonvertragssysteme und sollten die Aktionen ausführen, die für den Abschluss eines Mobilfunkvertrages in dem jeweiligen System erforderlich waren, wie z. B. das Auswählen von Vertragsoptionen oder die Angabe von fiktiven persönlichen Daten. Sie wurden vorher instruiert, dass sie mit einem fiktiven Schein-System arbeiten werden und demnach keinen rechtsgültigen Vertrag abschließen würden. Anschließend bewerteten die Teilnehmenden ihre Erfahrungen mit dem ihnen randomisiert zugeordneten webbasierten Mobiltelefonvertragssystem anhand des ISONORM 9241/10 und einer deutschen Übersetzung der positiv formulierten System Usability Scale, der Usability Metric for User Experience und der Adjective Rating Scale. Die Fragebogen wurden jedem Teilnehmenden in randomisierter Reihenfolge präsentiert.

Um den Unterschied in der wahrgenommenen Benutzbarkeit zwischen den beiden Systemen zu bewerten, wurden Korrelationen zwischen der Variable, die den Typ des Systems kodiert (gut gestaltetes System vs. schlecht gestaltetes System) einerseits und den Benutzbarkeitsbewertungen (Summenwerte) für jeden der vier Fragebogen der wahrgenommenen Benutzbarkeit andererseits berechnet. Diese Korrelationen stellen die Stichproben-Effektgrößenmaße dar und geben an, wie groß der Unterschied in der wahrgenommenen Benutzbarkeit zwischen dem gut und dem schlecht gestalteten System ist bzw. wie gut der jeweilige Fragebogen diesen Unterschied reflektiert. Um festzustellen, ob sich die vier verbalen Fragebogen signifikant darin unterscheiden, wie gut sie den Unterschied in der wahrgenommenen Benutzbarkeit zwischen den beiden Systemen reflektieren, wurden diese Korrelationen mit dem Steiger Test für Unterschiede zwischen abhängigen und überlappenden Korrelationen (Steiger, 1980) statistisch miteinander verglichen. Die Korrelationen sind abhängig, weil alle Teilnehmenden alle Fragebogen ausfüllten; die Korrelationen sind überlappend, weil eine Variable – die Variable, die den Systemtyp kodiert – Teil beider zu vergleichender Korrelationen ist.

Die statistische Auswertung zeigte, dass das kürzeste und damit effizienteste Instrument, die Adjective Rating Scale, mit nur einem Item den Unterschied in der wahrgenommenen Benutzbarkeit zwischen dem gut gestalteten System und dem schlecht gestalteten System deskriptiv besser reflektierte als alle längeren

Fragebogen. Die Ein-Item Adjective Rating Scale reflektierte diesen Unterschied in der wahrgenommenen Benutzbarkeit sogar signifikant besser als die 4-Item Usability Metric for User Experience und der 35-Item ISONORM 9241/10. Der Unterschied zwischen der Ein-Item Adjective Rating Scale und der 10-Item System Usability Scale war hingegen nicht signifikant. Die oft als Industriestandard bezeichnete 10-Item System Usability Scale reflektierte den Unterschied in der wahrgenommenen Benutzbarkeit signifikant besser als die 4-Item Usability Metric for User Experience. Außerdem zeigte sich ein deskriptiver Vorteil der 10-Item System Usability Scale im Vergleich zum 35-Item ISONORM 9241/10, allerdings wird dieser aufgrund der Kontrolle der Alphafehler-Kumulierung als nicht signifikant eingestuft. Zuletzt konnte kein signifikanter Unterschied zwischen dem 35-Item ISONORM 9241/10 und der 4-Item Usability Metric for User Experience darin gefunden werden, Unterschiede in der wahrgenommenen Benutzbarkeit zu reflektieren. Der ISONORM 9241/10 bietet auf Basis dieser Ergebnisse keine bessere Unterscheidbarkeit zwischen den zwei Systemen als kürzere Fragebogen. Dies steht im Gegensatz zu den Erwartungen auf der Basis der einschlägigen Literatur zur Messung psychologischer Konstrukte mit Fragebogen, nämlich dass Fragebogen mit mehr Items validere Messungen erlauben als Fragebogen mit weniger Items. Bevor jedoch weitere Schlussfolgerungen über die relative Nützlichkeit der jeweiligen Fragebogen gezogen werden konnten, war es notwendig, die Robustheit dieses Befundmusters mit einer konzeptuellen Replikation von Experiment 1a zu testen.

Experiment 1b

Die Materialien, die Vorgehensweise und das Design waren in Experiment 1b identisch zu Experiment 1a, abgesehen von folgenden Ausnahmen: Es wurde auf das Testen des ISONORM 9241/10 verzichtet, da er in Experiment 1a trotz seiner hohen Anzahl von Items keine signifikant bessere Performanz als alle kürzeren Fragebogen zeigte und somit am wenigsten effizient schien. Statt nicht-validierter deutscher Übersetzungen wurden zudem die validierten englischen Originalversionen der positiv formulierten System Usability Scale, der Usability Metric for User Experience und der Adjective Rating Scale verwendet, die von in Großbritannien lebenden Teilnehmenden bearbeitet und ausgefüllt wurden. Entsprechend wurden auch die Benutzeroberflächen der webbasierten Mobiltelefonvertragssysteme ins Englische übersetzt. Zusätzlich wurden die

Benutzeroberflächen der beiden Systeme so verändert, dass sich der Unterschied in ihrer wahrgenommenen Benutzbarkeit weiter vergrößert mit dem Ziel, einen „großen“ Unterschied in der Effektgröße im Sinne der von Cohen (1988) vorgeschlagenen Konventionen zu erreichen. So wurde beispielsweise der Text-Hintergrund-Kontrast verringert und rechts- und linksbündiger Text vermischt, um das schlecht gestaltete System noch weniger benutzbar zu gestalten. Wie in Experiment 1a arbeiteten die Teilnehmenden in Experiment 1b mit einem der beiden webbasierten Mobiltelefonvertragssysteme. Anschließend sollten sie ihre Erfahrungen mit dem jeweiligen System anhand der positiv formulierten System Usability Scale, der Usability Metric for User Experience und der Adjective Rating Scale bewerten, die in randomisierter Reihenfolge allen Teilnehmenden vorgelegt wurden.

Der statistische Vergleich der drei Fragebogen – System Usability Scale, Usability Metric for User Experience und Adjective Rating Scale – wurde wieder mit Hilfe des Steiger Tests für Unterschiede zwischen abhängigen und überlappenden Korrelationen (Steiger, 1980) durchgeführt. Parallel zu Experiment 1a war die Adjective Rating Scale in Experiment 1b den längeren Fragebogen bezüglich des Reflektierens des Unterschiedes in wahrgenommener Benutzbarkeit zwischen den zwei Systemen deskriptiv überlegen. Im Gegensatz zu Experiment 1a reflektierte die Ein-Item Adjective Rating Scale den Unterschied in der wahrgenommenen Benutzbarkeit zwischen dem gut und dem schlecht gestalteten System in Experiment 1b sogar signifikant besser als die 10-Item System Usability Scale, während jedoch der Unterschied zur 4-Item Usability Metric for User Experience nicht signifikant war. Der in Experiment 1a signifikante Unterschied zwischen der 10-Item System Usability Scale und der 4-Item Usability Metric for User Experience ließ sich in Experiment 1b nicht replizieren. Es wurde kein signifikanter Unterschied zwischen den beiden Fragebogen gefunden.

Diskussion

In den Experimenten 1a und 1b zeigte sich, dass der im Einsatz kostengünstige und effizienteste verbale Fragebogen zur Messung der wahrgenommenen Benutzbarkeit – die Ein-Item Adjective Rating Scale – die Unterschiede in der wahrgenommenen Benutzbarkeit zweier Systeme mindestens genauso gut und in einigen Fällen sogar

signifikant besser als etablierte längere Fragebogen reflektierte. Damit wird die Annahme unterstützt, dass die verbale Ein-Item Adjective Rating Scale für die Bewertung der wahrgenommenen Benutzbarkeit geeignet sein könnte.

Neben dem, was in der wissenschaftlichen Literatur zur Messung psychologischer Konstrukte mit Fragebogen als Mehrheitsmeinung gelten kann, findet man durchaus Überlegungen dazu, dass und unter welchen Umständen Ein-Item-Fragebogenmaße psychologischer Konstrukte angemessen sein könnten. Rossiter (2002) beispielsweise hat vorgeschlagen, dass solche Messungen generell dann angemessen sind, wenn das zu messende psychologische Konstrukt und seine Attribute als sogenannte konkrete und singuläre Entitäten (Objekte) angesehen werden können. Die Bedingungen dafür lauten, dass (1) alle Bewertenden verstehen, welche Entität, also welches Objekt, bewertet wird und (2), dass das, was bewertet wird, relativ homogen ist. Die Erfüllung der ersten Bedingung scheint in diesen hier betrachteten Experimenten plausibel zu sein, da den Teilnehmenden nur ein Item zur Bewertung eines einfachen psychologischen Konstruktes – wahrgenommene Benutzbarkeit – präsentiert wurde (Bangor et al., 2009). Um zu prüfen, ob auch die zweite Bedingung, die Homogenität des psychologischen Konstruktes der wahrgenommenen Benutzbarkeit, erfüllt ist, wurde geprüft, ob die 10-Item System Usability Scale und die 4-Item Usability Metric for User Experience als eindimensionale Maße betrachtet werden können. Dies basierte auf der Annahme, dass eine einfaktorielle Lösung als Ergebnis einer explorativen Faktorenanalyse auf die mögliche Eindimensionalität der beiden Maße und damit des gemessenen psychologischen Konstrukts im Allgemeinen hinweisen würde. Genau genommen ist es eine notwendige Bedingung für die Homogenität eines psychologischen Konstrukts, dass sich für die Maße, die das psychologische Konstrukt erfassen sollen, in einer explorativen Faktorenanalyse Eindimensionalität zeigt. Um die Eindimensionalitätsannahme zu prüfen, wurde eine Parallelanalyse (Horn, 1965) mit den Daten der vorliegenden zwei Experimente, unter Verwendung der Hauptkomponentenextraktion und Beibehaltung aller Faktoren mit Eigenwerten, die größer als das 95. Perzentil der Referenzeigenwerte sind (Auerswald & Moshagen, 2019), durchgeführt. Diese Analyse ergab, dass sowohl die System Usability Scale als auch die Usability Metric for User Experience im Wesentlichen eindimensional behandelt werden können (vergleiche Gräve & Buchner, 2024). Diese Ergebnisse lassen die Schlussfolgerung zu, dass sie einen Indikator für die Homogenität des

psychologischen Konstruktes der wahrgenommenen Benutzbarkeit im Sinne von Rossiter (2002) liefern. Im Kern passen die vorliegenden Ergebnisse also in den von Rossiter (2002) vorgeschlagenen theoretischen Rahmen.

Auf Basis der Ergebnisse der Experimente 1a und 1b lässt sich zusammenfassend schließen, dass die verbale Adjective Rating Scale mit nur einem Item als eine valide und effiziente Alternative zu mehrstufigen verbalen Fragebogen zur Messung wahrgenommener Benutzbarkeit angesehen werden kann. Um die Robustheit dieser Befunde bezüglich der beiden verbalen Fragebogen – System Usability Scale und Adjective Rating Scale – zu prüfen und um zu prüfen, ob ihre piktoralen Alternativen ein vergleichbares Ergebnismuster aufweisen, dienten die folgenden Experimente 2a und 2b.

Vergleich von verbalen und piktoralen Ein-Item- und 10-Item-Fragebogen zur Messung der wahrgenommenen Benutzbarkeit

In der zweiten Testreihe wurde zuerst geprüft, ob die Befunde bezüglich der beiden verbalen Fragebogen – der 10-Item System Usability Scale und der Ein-Item Adjective Rating Scale – replizierbar sind. Verbale Fragebogen können jedoch den Nachteil haben, dass Personen, welche die verwendete Sprache nur begrenzt beherrschen, einzelne Items falsch interpretieren und daraus resultierend unbeabsichtigte Antworten geben (Bradley & Lang, 1994; Kunin, 1955). Als eine Alternative zu den verbalen Fragebogen der wahrgenommenen Benutzbarkeit wurden piktorale Fragebogen entwickelt, da sie generell als leicht verständlich und intuitiv zu beantworten gelten (Bradley & Lang, 1994). Basierend auf der Tatsache, dass der verbale Ein-Item-Fragebogen in Experiment 1a und 1b den Unterschied in der wahrgenommenen Benutzbarkeit sehr gut reflektiert hat, wurde in der folgenden Testreihe getestet, ob ein ähnliches Befundmuster für eine piktorale Ein-Item-Alternative zur Messung der wahrgenommenen Benutzbarkeit gezeigt werden kann. Darüber hinaus wurde geprüft, ob die piktoralen Alternativen den Unterschied in der wahrgenommenen Benutzbarkeit zwischen dem gut gestalteten System und dem schlecht gestalteten System genauso gut reflektieren wie ihre verbalen Gegenstücke.

Experiment 2a

In Experiment 2a wurden vier Fragebogen zur Messung wahrgenommener Benutzbarkeit experimentell verglichen, zwei verbale und zwei piktorale Fragebogen. Es wurde getestet, wie gut die verbalen Fragebogen – die 10-Item positiv formulierte System Usability Scale (Sauro & Lewis, 2011) und die Ein-Item Adjective Rating Scale (Bangor et al., 2009) – und deren piktorale Alternativen – die 10-Item Pictorial System Usability Scale (Baumgartner et al., 2019a) und die Ein-Item Pictorial Single-Item Usability Scale (Baumgartner et al., 2019b) – den Unterschied in der wahrgenommenen Benutzbarkeit zwischen zwei Systemen, die sich in ihrer Benutzbarkeit unterscheiden, reflektieren. Die beiden webbasierten Mobiltelefonvertragssysteme aus der ersten Testreihe (Experiment 1a und 1b) wurden dahingehend verändert, dass den Teilnehmenden nun auf Basis ihrer angegebenen Wünsche (z. B. zusätzliches Mobiltelefon oder eine zweite SIM-Karte) zwei Tarifmöglichkeiten angeboten wurden, die sich basierend auf den angegebenen Wünschen veränderten. In den Experimenten 1a und 1b wurden nur zwei gleichbleibende Tarifmöglichkeiten präsentiert mit der Option, zusätzliche Angebote hinzu zu buchen. Durch diese Änderung ergibt sich für die Experimente 2a und 2b im gut gestalteten System die Möglichkeit, unter Verwendung des Zurück-Knopfes die eigenen Wünsche noch einmal zu ändern und sich neue Preise vorschlagen zu lassen, was die Benutzbarkeit des Systems erhöhen sollte. Im schlecht gestalteten System gab es die Möglichkeit der Änderung nicht, da bewusst kein Zurück-Knopf angezeigt wurde. Da viele Teilnehmende in der ersten Testreihe das Experiment an der Stelle abgebrochen haben, an der eine fiktive Fehlermeldung präsentiert wurde, wurde für die folgenden Experimente in beiden Systemen von Fehlermeldungen abgesehen. Um mögliche Abbrüche von Teilnehmenden aufgrund des schlecht lesbaren Textes für die folgenden Experimente zu vermeiden, wurde die extrem schlechte Lesbarkeit des Textes im schlecht gestalteten System für die Experimente 2a und 2b im Vergleich zu Experiment 1b leicht verbessert– graue statt dunkelblaue Schrift auf schwarzem Hintergrund.

Nachdem die in Großbritannien lebenden Teilnehmenden mit einem der beiden webbasierten Mobiltelefonvertragssysteme interagiert hatten, bewerteten sie ihre Erfahrungen mit dem System entweder anhand der englischen Versionen der zwei verbalen Fragebogen der wahrgenommenen Benutzbarkeit – der positiv

formulierten System Usability Scale und der Adjective Rating Scale – oder anhand der zwei piktoralen Fragebogen der wahrgenommenen Benutzbarkeit – der Pictorial System Usability Scale und der Pictorial Single-Item Usability Scale.

Um jeweils die beiden verbalen Fragebogen miteinander sowie die beiden piktoralen Fragebogen miteinander statistisch zu vergleichen, wurde der Steiger Test für Unterschiede zwischen abhängigen und überlappenden Korrelationen (Steiger, 1980) durchgeführt. Die statistische Auswertung zeigte, dass es hinsichtlich ihrer Fähigkeit, den Unterschied in der wahrgenommenen Benutzbarkeit zwischen dem gut gestalteten System und dem schlecht gestalteten System zu reflektieren, keinen signifikanten Unterschied zwischen der verbalen Ein-Item Adjective Rating Scale und der verbalen 10-Item System Usability Scale gab. Deskriptiv war die Adjective Rating Scale jedoch überlegen. Das gleiche Befundmuster wurde für die beiden piktoralen Fragebogen gezeigt: Wieder war die Messung der wahrgenommenen Benutzbarkeit mit dem Ein-Item-Fragebogen, der Pictorial Single-Item Usability Scale, statistisch genauso gut wie die Messung mit deren Mehr-Item-Alternative, der 10-Item Pictorial System Usability Scale – kein signifikanter Unterschied zwischen den beiden Fragebogen. Deskriptiv war der Ein-Item-Fragebogen sogar besser als der 10-Item-Fragebogen. Diese Ergebnisse stützen die Befunde aus den Experimenten 1a und 1b, die zeigen, dass die verbale Ein-Item Adjective Rating Scale die Unterschiede in der wahrgenommenen Benutzbarkeit zweier Systeme mindestens genauso gut wie längere verbale Fragebogen reflektierte. Zusätzlich weisen die Ergebnisse darauf hin, dass eine ähnliche Schlussfolgerung für die piktorale Ein-Item-Alternative zur Messung wahrgenommener Benutzbarkeit, die Pictorial Single-Item Usability Scale, gelten könnte, da sie den Unterschied in der wahrgenommenen Benutzbarkeit mindestens genauso gut reflektierte wie die längere 10-Item Pictorial System Usability Scale.

Um schließlich zu testen, ob es einen Unterschied zwischen den verbalen und den piktoralen Fragebogen der wahrgenommenen Benutzbarkeit gibt, wurden die verbalen Fragebogen – die System Usability Scale und die Adjective Rating Scale – jeweils mit ihrem piktoralen Gegenstück – die Pictorial System Usability Scale und die Pictorial Single-Item Usability Scale – unter Verwendung des zweiseitigen Tests für unabhängige Korrelation (Fisher, 1925) verglichen. Die Korrelationen sind unabhängig, da die verbalen und die piktoralen Fragebogen von disjunkten

Gruppen von Teilnehmenden ausgefüllt wurden. Die statistische Auswertung zeigte, dass die verbale 10-Item System Usability Scale den Unterschied in der wahrgenommenen Benutzbarkeit zwischen den beiden Systemen signifikant besser als die piktorale 10-Item Pictorial System Usability Scale reflektierte. Dasselbe Ergebnis zeigte sich für die verbale Ein-Item Adjective Rating Scale, die den Unterschied in der wahrgenommenen Benutzbarkeit zwischen den beiden Systemen signifikant besser als die Ein-Item Pictorial Single-Item Usability Scale reflektierte. Um die Robustheit dieser Befundmuster zu testen, wurde eine konzeptuelle Replikation von Experiment 2a durchgeführt.

Experiment 2b

Die Materialien, die Vorgehensweise und das Design waren in Experiment 2b identisch zu Experiment 2a, abgesehen von einer Änderung: Statt der rein positiv formulierten System Usability Scale (Sauro & Lewis, 2011) wurde die ursprüngliche System Usability Scale (Brooke, 1996) mit einem Wechsel aus positiv und negativ formulierten Items verwendet. Allerdings wurde die Version der ursprünglichen System Usability Scale verwendet, in der in Item 8 das Wort „*awkward*“ anstatt des ursprünglichen Wortes „*cumbersome*“ benutzt wird (Bangor et al., 2008; Finstad, 2006). Dies basiert auf der Grundlage, dass die positiv formulierte System Usability Scale zwar aufgrund ihrer einheitlich positiv formulierten Items für nicht überwachte Studien-, wie die Online-Umgebung in den Experimenten dieser Dissertation empfohlen wird (Kortum et al., 2021; Sauro & Lewis, 2011), es jedoch zu klären bleibt, inwieweit die Verwendung der ursprünglichen System Usability Scale mit einem Wechsel zwischen positiv und negativ formulierten Items in so einem Online-Kontext tatsächlich zu Problemen führen kann. Darüber hinaus wird durch die Verwendung der ursprünglichen System Usability Scale eine bessere Vergleichbarkeit zwischen der verbalen und der piktoralen System Usability Scale geschaffen, da bei der Erstellung der Pictorial System Usability Scale die abwechselnd positiven und negativen Formulierungen der ursprünglichen System Usability Scale bildlich umgesetzt wurden. Wie auch in Experiment 2a bewerteten die Teilnehmenden nach einer kurzen Arbeitsphase ihre Erfahrung mit einem der beiden webbasierten Mobiltelefonvertragssysteme. Dies erfolgte in Experiment 2b anhand von zwei verbalen Fragebogen der wahrgenommenen Benutzbarkeit – der ursprünglichen System Usability Scale und der Adjective Rating Scale – oder anhand

von zwei piktoralen Fragebogen der wahrgenommenen Benutzbarkeit – der Pictorial System Usability Scale und der Pictorial Single-Item Usability Scale.

Der statistische Vergleich der beiden verbalen Fragebogen – der ursprünglichen System Usability Scale und der Adjective Rating Scale – mithilfe des Steiger Tests (Steiger, 1980) zeigte, dass die Ein-Item Adjective Rating Scale den Unterschied in der wahrgenommenen Benutzbarkeit beider Systeme signifikant besser reflektierte als die 10-Item ursprüngliche System Usability Scale. Zwischen den beiden piktoralen Fragebogen – der Pictorial Single-Item Usability Scale und der Pictorial System Usability Scale – konnte kein signifikanter Unterschied gefunden werden. Diese Ergebnisse stützen das Befundmuster aus Experiment 2a: Ein-Item-Fragebogen, verbal oder piktoral, reflektierten die Unterschiede in der wahrgenommenen Benutzbarkeit zweier Systeme mindestens genauso gut wie längere Fragebogen. Der statistische Vergleich der verbalen und der piktoralen Fragebogen der wahrgenommenen Benutzbarkeit wurde unter Verwendung des zweiseitigen Tests für unabhängige Korrelation (Fisher, 1925) durchgeführt. Die verbale ursprüngliche System Usability Scale wurde mit ihrem piktoralen Gegenstück der Pictorial System Usability Scale und die verbale Ein-Item Adjective Rating Scale mit der Pictorial Single-Item Usability Scale verglichen. Es wurde kein signifikanter Unterschied zwischen der verbalen ursprünglichen System Usability Scale und der Pictorial System Usability Scale darin gefunden, den Unterschied in der wahrgenommenen Benutzbarkeit zwischen dem gut gestalteten System und dem schlecht gestalteten System zu reflektieren. Auch der statistische Vergleich der verbalen Adjective Rating Scale und der Pictorial Single-Item Usability Scale zeigte keinen signifikanten Unterschied zwischen den beiden Ein-Item-Fragebogen darin, den Unterschied in der wahrgenommenen Benutzbarkeit zu reflektieren. Im Gegensatz zu Experiment 2a konnte somit kein Vorteil der verbalen Fragebogen gegenüber den piktoralen Fragebogen beobachtet werden.

Diskussion

Die Ergebnisse der Experimente 2a und 2b zeigen, dass die Befunde bezüglich der verbalen Fragebogen aus den Experimenten 1a und 1b der vorliegenden Dissertation repliziert werden konnten. Die verbale Ein-Item Adjective Rating Scale zur Messung der wahrgenommenen Benutzbarkeit reflektierte die Unterschiede in der

wahrgenommenen Benutzbarkeit zweier Systeme genauso gut wie die positiv formulierte 10-Item System Usability Scale und sogar signifikant besser als die ursprüngliche System Usability Scale. Basierend auf diesen Befunden kann angenommen werden, dass die Ein-Item Adjective Rating Scale eine valide und effiziente Alternative zu längeren Fragebogen zur Messung der wahrgenommenen Benutzbarkeit darstellt. Eine parallele Annahme kann für die piktorale Ein-Item-Alternative, die Pictorial Single-Item Usability Scale, getroffen werden, da für diese ein ähnliches Befundmuster gezeigt werden konnte. Die Pictorial Single-Item Usability Scale reflektierte in den Experimenten 2a und 2b den Unterschied in der wahrgenommenen Benutzbarkeit zwischen zwei Systemen mindestens genauso gut wie die längere 10-Item Pictorial System Usability Scale. Die Pictorial Single-Item Usability Scale kann daher als eine einfach durchzuführende und effiziente Alternative zur 10-Item Pictorial System Usability Scale betrachtet werden. In den Experimenten 2a und 2b zeigten sich jedoch gemischte Ergebnisse hinsichtlich des Vergleichs der verbalen Fragebogen und ihren piktoralen Pendants. Der in Experiment 2a beobachtete Vorteil der verbalen Fragebogen gegenüber den piktoralen Fragebogen konnte in Experiment 2b nicht repliziert werden. Diese Ergebnisse deuten darauf hin, dass piktorale Fragebogen im Vergleich zu den verbalen Fragebogen möglicherweise eine geringfügig reduzierte Validität aufweisen. Diese Verringerung scheint jedoch so gering zu sein, dass sie kein starkes Argument gegen den Einsatz piktoraler Fragebogen zur Messung der wahrgenommenen Benutzbarkeit darstellt.

Im Hinblick auf die verbalen Fragebogen wurden in den Experimenten 2a und 2b zwei verschiedene Versionen der System Usability Scale verwendet, in Experiment 2a die positiv formulierte Version der System Usability Scale (Sauro & Lewis, 2011) und in Experiment 2b die ursprüngliche System Usability Scale (Brooke, 1996). In der vorliegenden Testreihe lieferten die beiden Versionen der System Usability Scale im Wesentlichen gleiche Ergebnisse. Dies stützt die Schlussfolgerung von Sauro und Lewis (2011) und Kortum et al. (2021), dass beide Versionen der System Usability Scale bedenkenlos verwendet werden können, da ihre Ergebnisse und die daraus resultierenden Entscheidungen vergleichbar sind.

Zusammenfassend lässt sich nach den Experimenten 2a und 2b sagen, dass sowohl der verbale (Adjective Rating Scale) als auch der piktorale (Pictorial Single-Item

Usability Scale) Ein-Item-Fragebogen den Unterschied in der wahrgenommenen Benutzbarkeit zwischen einem gut und einem schlecht gestalteten System mindestens genauso gut reflektiert wie sein 10-Item-Pendant, die verbale System Usability Scale oder die Pictorial System Usability Scale.

Allgemeine Diskussion

In der vorliegenden Dissertation wurden verbale und piktorale standardisierte Fragebogen zur Messung der wahrgenommenen Benutzbarkeit, die sich in ihrer Länge und Komplexität unterscheiden, experimentell verglichen. Genauer gesagt wurde geprüft, wie gut diese Fragebogen den Unterschied in der wahrgenommenen Benutzbarkeit zwischen zwei absichtlich unterschiedlich gestalteten Systemen reflektieren. Um zu testen, ob das psychologische Konstrukt der wahrgenommenen Benutzbarkeit besser durch Fragebogen mit einer größeren oder mit einer kleineren Anzahl von Items gemessen werden kann, wurden im Rahmen der ersten Testreihe der vorliegenden Dissertation in Experiment 1a und 1b vier verbale Fragebogen mit unterschiedlicher Länge und Komplexität experimentell verglichen. Das Hauptergebnis der ersten Testreihe lautet, dass der kürzeste der hier betrachteten verbalen Fragebogen der wahrgenommenen Benutzbarkeit, die Ein-Item Adjective Rating Scale, die Unterschiede in der wahrgenommenen Benutzbarkeit zwischen zwei Systemen mindestens genauso gut und in einigen Fällen sogar signifikant besser als etablierte, längere Fragebogen reflektierte. In den Experimenten 2a und 2b wurde die Robustheit dieses Ergebnismusters in Bezug auf die beiden verbalen Fragebogen – System Usability Scale und Adjective Rating Scale – getestet. Die Befunde aus den Experimenten 1a und 1b konnten repliziert werden, da die verbale Ein-Item Adjective Rating Scale die Unterschiede in der wahrgenommenen Benutzbarkeit beider Systeme mindestens genauso gut (Experiment 2a) oder sogar signifikant besser (Experiment 2b) als die längere 10-Item System Usability Scale reflektierte. Darüber hinaus wurde in den Experimenten 2a und 2b getestet, ob piktorale Alternativen der beiden verbalen Fragebogen – die Pictorial System Usability Scale und die Pictorial Single-Item Usability Scale – ein vergleichbares Ergebnismuster aufweisen. Auch die piktorale Ein-Item-Alternative, die Pictorial Single-Item Usability Scale, reflektierte in den Experimenten 2a und 2b den Unterschied in der wahrgenommenen Benutzbarkeit zwischen den beiden Systemen mindestens genauso gut wie die längere 10-Item Pictorial System Usability Scale.

Angesichts der Ergebnisse der vier durchgeführten Experimente und in Anbetracht der wirtschaftlichen und pragmatischen Vorteile von Ein-Item-Messungen kann geschlussfolgert werden, dass sowohl die verbale Ein-Item Adjective Rating Scale als auch die piktorale Ein-Item Pictorial Single-Item Usability Scale als valide und effiziente Alternativen zu ihren Multi-Item-Pendants der wahrgenommenen Benutzbarkeit angesehen werden können.

Auf Grundlage der vorliegenden Ergebnisse dieser Dissertation lässt sich über eine mögliche Erklärung für den Erfolg der Ein-Item-Fragebogen spekulieren. Da der Text des verbalen und die bildliche Darstellung des piktoralen Ein-Item-Fragebogens relativ allgemein gehalten und somit interpretationsoffen sind, könnten sich die Benutzenden flexibel auf diejenige Determinante konzentriert haben, die in einem bestimmten Kontext die wahrgenommene Benutzbarkeit am stärksten bestimmt (wie z. B. die schlechte Lesbarkeit des Textes mit negativer Polarität in dem schlecht gestalteten System). Dies könnte den Unterschied zwischen den Systemen im Hinblick auf die wahrgenommene Benutzbarkeits-Bewertung maximieren (vergleiche Gräve et al., 2024; Gräve & Buchner, 2024). Ob dies tatsächlich der Fall ist, sollte in einer unabhängigen Studie empirisch untersucht werde. Darüber hinaus könnte eine mögliche Erklärung für die schlechte Performanz längerer Fragebogen (wie des ISONORM 9241/10) sein, dass die Verwendung verbaler und piktoraler Multi-Item-Fragebogen der wahrgenommenen Benutzbarkeit dazu führen, dass Teilnehmende irrelevante oder unbekannte Aspekte einer Anwendung bewerten müssen. So haben Teilnehmende möglicherweise nach der Interaktion mit dem System keine Grundlage für die Beantwortung mancher verbaler oder piktoraler Items der längeren Fragebogen, auf die sie zurückgreifen könnten. Dies könnte als Konsequenz einen negativen Einfluss darauf haben, wie gut ein langer Fragebogen der wahrgenommenen Benutzbarkeit die Unterschiede in der wahrgenommenen Benutzbarkeit zwischen zwei Systemen reflektiert. Ob diese Erklärung tatsächlich zutrifft, sollte in einer unabhängigen Studie empirisch untersucht werden.

Die vorliegenden Ergebnisse passen sehr gut in den theoretischen Rahmen von Rossiter (2002), der als Voraussetzung für die Verwendung von Einzel-Items zur Messung eines psychologischen Konstruktes angibt, dass das zu messende psychologische Konstrukt und seine Attribute als konkrete und singuläre Entitäten

angesehen werden können. Im Falle der wahrgenommenen Benutzbarkeit trifft es vermutlich zu, dass (1) alle Bewertende verstehen, welche Entität bewertet wird, und, dass (2) das, was bewertet wird, relativ homogen ist (siehe Diskussion zu Experiment 1a und 1b, Bangor et al., 2009; Berkman & Karahoca, 2016; Kortum et al., 2021; Lewis, 2019; Lewis & Sauro, 2017; Sauro, 2018). Daraus lässt sich eine mögliche Heuristik ableiten, die besagt, dass ein Ein-Item-Maß der wahrgenommenen Benutzbarkeit dann für die Messung des psychologischen Konstruktes angemessen ist, wenn das zu evaluierende System so einfach ist, dass seine Benutzeroberfläche als konkret-singulär im Sinne von Rossiter (2002) angesehen werden kann (vergleiche Gräve & Buchner, 2024). Bei komplexeren Systemen könnten hingegen, wie von Bangor et al. (2009) beschrieben, verschiedene Teile der Benutzeroberfläche, wie beispielsweise die Hauptnavigation im Gegensatz zum Hilfesystem, unterschiedlich bewertet werden, sodass die Benutzeroberfläche nicht unbedingt als singulär angesehen werden könnte. Darüber hinaus wurde die Benutzbarkeit der in der vorliegenden Dissertation entworfenen webbasierten Mobiltelefonvertragssysteme primär auf der perzeptuellen Ebene manipuliert, indem beispielsweise die Manipulation des Text-Hintergrund-Kontrasts oder die Manipulation der räumlichen Anordnung der Dialogelemente variiert wurde. Innerhalb dieses Kontextes konnte gezeigt werden, dass die Ein-Item-Fragebogen die wahrgenommene Benutzbarkeit mindestens genauso gut – wenn nicht sogar besser – als ihre Multi-Item-Pendants messen. Werden jedoch Ein-Item-Fragebogen mit Multi-Item-Fragebogen der wahrgenommenen Benutzbarkeit unter Verwendung komplexerer Systeme mit facettenreicheren Manipulationen auf mehreren Ebenen – wie z. B. einer Manipulation der Menüstrukturen und der Verständlichkeit dieser Menüstrukturen – verglichen, könnten Multi-Item-Fragebogen einen Vorteil haben, da sie unterschiedliche Aspekte der wahrgenommenen Benutzbarkeit abfragen und dementsprechend zusätzliche Informationen liefern. Ein-Item-Fragebogen hingegen können nur einen Gesamteindruck der wahrgenommenen Benutzbarkeit erfassen und daher bei komplexeren Systemen die verschiedenen Dimensionen der Benutzbarkeit dieser Systeme möglicherweise nicht ausreichend abbilden. Um diese theoretischen Annahmen und die aufgestellte Heuristik zu prüfen, sollte zukünftig getestet werden, wie gut Ein-Item- und Multi-Item-Fragebogen den Unterschied in der

wahrgenommenen Benutzbarkeit zwischen zwei komplexen Systemen, die sich in ihrer Benutzbarkeit unterscheiden, reflektieren.

Wie bereits erwähnt, sind die verbalen und piktoralen Ein-Item-Fragebogen zur Messung der wahrgenommenen Benutzbarkeit jedoch nur innerhalb bestimmter Grenzen anwendbar. Sie können nur einen zusammenfassenden Eindruck von der wahrgenommenen Benutzbarkeit eines Systems vermitteln und somit nur Auskunft über den Grad der wahrgenommenen Benutzbarkeit eines Systems oder den Unterschied in der wahrgenommenen Benutzbarkeit zwischen Systemen geben. Dies hat Vorteile, wenn es darum geht, einfach und schnell einen Gesamteindruck über die wahrgenommene Benutzbarkeit des Systems zu erhalten. Bei einer formativen Evaluierung, bei der fortlaufend Bewertungen eines Produktes während der Entwicklungsphase erhoben werden, um das Produkt iterativ zu verbessern, ist jedoch ein Ein-Item-Fragebogen nur von begrenztem Nutzen, da die Gründe für die Bewertung unklar bleiben. Hier könnte es Vorteile haben Multi-Item-Fragebogen zu verwenden, da sie zusätzliche Informationen liefern können, die zur Identifikation spezifischer Probleme (Baumgartner et al., 2019b) oder für die Lokalisierung von Benutzbarkeits-Schwächen (Baumgartner et al., 2019a) hilfreich sein könnten. Basierend auf diesen Überlegungen sollten die verbalen und die piktoralen Ein-Item-Fragebogen der wahrgenommenen Benutzbarkeit nur als effiziente Alternativen zu den Gesamtwerten ihrer Multi-Item-Pendants betrachtet werden.

Eine offensichtliche Einschränkung der vorliegenden Experimente im Rahmen dieser Dissertation besteht darin, dass jeder Teilnehmende die wahrgenommene Benutzbarkeit des Systems mit mindestens zwei Fragebogen und bis zu maximal vier Fragebogen bewertete, wenn auch in einer für jeden Teilnehmenden randomisierten Reihenfolge. Dieses Vorgehen ist zwar effizient, birgt aber auch die Gefahr von Übertragungseffekten in dem Sinne, dass das Ausfüllen früherer Fragebogen die Art und Weise beeinflussen kann, wie spätere Fragebogen ausgefüllt werden (vergleiche Gräve & Buchner, 2024). Um Übertragungseffekte zu vermeiden, sollte in zukünftigen Studien eine Person nur einen Fragebogen zur Messung wahrgenommener Benutzbarkeit ausfüllen. Jedoch ist dann der experimentelle Vergleich von Fragebogen zur Messung wahrgenommener Benutzbarkeit realistischerweise auf zwei Fragebogen begrenzt, da bei dem Vergleich von mehr als zwei Fragebogen die Anforderungen an die Stichprobengröße schnell sehr groß

werden. Ob und inwieweit diese methodische Änderung einen Einfluss auf das hier vorliegende Befundmuster bezüglich des Vergleichs von Ein-Item- und Multi-Item-Fragebogen wahrgenommener Benutzbarkeit hat, könnte für die zukünftige Forschung von Interesse sein.

Wie bereits in der Diskussion der zweiten Testreihe kurz beschrieben, konnte kein konsistenter Vorteil der verbalen Fragebogen gegenüber den piktoralen Fragebogen gezeigt werden. In Experiment 2a reflektierten die verbalen Fragebogen den Unterschied in der wahrgenommenen Benutzbarkeit signifikant besser als die piktoralen Fragebogen, jedoch konnte dies in Experiment 2b nicht repliziert werden. Daraus lässt sich schließen, dass piktorale Fragebogen im Vergleich zu den verbalen Fragebogen möglicherweise eine geringfügig reduzierte Validität aufweisen. Diese Verringerung scheint jedoch so gering zu sein, dass sie kein starkes Argument gegen den Einsatz piktoraler Fragebogen zur Messung der wahrgenommenen Benutzbarkeit darstellt (vergleiche Gräve et al., 2024). Vor allem in Situationen, in denen piktorale Instrumente spezifische Vorteile haben, wie beispielweise beim Einbeziehen von Nutzenden mit unterschiedlichen Sprachkenntnissen, Nutzenden mit Legasthenie oder Nutzenden mit niedrigem Bildungsniveau (Baumgartner et al., 2019b), scheint es auf Basis der vorliegenden Ergebnisse gerechtfertigt zu sein, piktorale Fragebogen einzusetzen. In Situationen, in denen die spezifischen Vorteile von piktoralen Fragebogen weniger wichtig sind, beispielsweise wenn die Sprachkenntnisse der Befragten keine Rolle spielen, könnte die Verwendung von verbalen Fragebogen angemessen sein, da sie möglicherweise die Unterschiede in der wahrgenommenen Benutzbarkeit etwas besser reflektieren können als ihre piktoralen Pendants (Experiment 2a). In zukünftigen Studien könnten die Fragebogen mit Nutzergruppen, die voraussichtlich besonders von piktoralen Fragebogen profitieren, explizit getestet werden. Dadurch könnte der theoretische Nutzen der piktoralen Fragebogen in Situationen, in denen piktorale Instrumente spezifische Vorteile haben, experimentell überprüft werden (Baumgartner et al., 2019b).

Abschließend gilt es zu erwähnen, dass die Ergebnisse der vorliegenden Dissertation von hoher praktischer Relevanz sind. Sie können wertvolle Impulse für Firmen der Softwareentwicklung liefern. Da die verbalen und piktoralen Ein-Item-Fragebogen der wahrgenommenen Benutzbarkeit auf Basis dieser Ergebnisse als valide

Alternativen zu ihren Multi-Items-Pendants angesehen werden können, werden sie für die Messung von Gesamtunterschieden in der wahrgenommenen Benutzbarkeit empfohlen. Bei Verwendung von Ein-Item-Fragebogen haben Fachleute aus der Praxis einen geringen Aufwand bei der Erstellung und Administration des Fragebogens und auch die Bearbeitungszeit der Befragten reduziert sich im Vergleich zu einem längeren Fragebogen (Pomeroy et al., 2001). So kann in Benutzbarkeitstestungen effizienter eine große Stichprobe an Befragten erhoben werden. Dies ist von Vorteil, da eine ansteigende Größe der verwendeten Stichprobe einen positiven Einfluss auf die Repräsentativität der Stichprobe sowie die Teststärke der Analyse und so insgesamt auf die Genauigkeit der Ergebnisse haben kann. Darüber hinaus könnte es eine Möglichkeit geben, Ein-Item-Fragebogen auch für den Vergleich von iterativen Versionen desselben Systems einzusetzen, um beispielsweise mehrere Zyklen in der Entwicklung der Software bis zum Endprodukt zu begleiten. Dies kann am Ende zu einer höheren Investitionsrendite führen (Marcus, 2005). Für die verbale System Usability Scale konnte bereits gezeigt werden, dass sie Änderungen der wahrgenommenen Benutzbarkeit während eines iterativen Designzyklus messen kann (Bangor et al., 2008). Ob jedoch auch die kostengünstigeren und effizienteren Ein-Item-Alternativen – die verbale Adjective Rating Scale und die Pictorial Single-Item Usability Scale – den Fortschritt (oder Rückschritt) in der wahrgenommenen Benutzbarkeit messen können, während ein System einen iterativen Designzyklus durchläuft, sollte in zukünftigen Studien überprüft werden.

Fazit

Zusammenfassend konnte in der vorliegenden Dissertation in allen vier Experimenten gezeigt werden, dass der kürzeste und kostengünstigste der betrachteten verbalen Fragebogen zur Messung der wahrgenommenen Benutzbarkeit, die verbale Ein-Item Adjective Rating Scale, den Unterschied in der wahrgenommenen Benutzbarkeit zwischen zwei Systemen mindestens genauso gut und in einigen Fällen sogar signifikant besser als etablierte, längere Multi-Item-Fragebogen reflektiert. Auch für die piktorale Ein-Item-Alternative, die Pictorial Single-Item Usability Scale, konnte gezeigt werden, dass diese den Unterschied in der wahrgenommenen Benutzbarkeit zwischen den beiden Systemen mindestens genauso gut reflektiert wie die längere 10-Item Pictorial System Usability Scale.

Angesichts dieser Ergebnisse und in Anbetracht des wirtschaftlichen Vorteils von Ein-Item-Fragebogen können sowohl die verbale Ein-Item Adjective Rating Scale als auch die Pictorial Single-Item Usability Scale als valide und effiziente Maße für die wahrgenommene Benutzbarkeit empfohlen werden.

Literatur

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, *58*(2), 218-227. https://doi.org/10.2501/JAR-2017-001

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468-491. https://doi.org/10.1037/met0000200

Bangor, A., Kortum, P., & Miller, J. (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

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114-123.

Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., & Sonderegger, A. (2019a). Pictorial System Usability Scale (P-SUS): Developing an instrument for measuring perceived usability. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK*. https://doi.org/10.1145/3290605.3300299

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019b). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies*, *122*, 78-89. https://doi.org/10.1016/j.ijhcs.2018.08.008

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175-184. https://doi.org/10.1509/jmkr.44.2.175

Bergkvist, L., & Rossiter, J. R. (2009). Tailor-made single-item measures of doubly concrete constructs. *International Journal of Advertising: The Quarterly Review of Marketing Communications*, *28*(4), 607-621. https://doi.org/10.2501/S0265048709200783

Berkman, M. I., & Karahoca, D. (2016). Re-assessing the Usability Metric for User Experience (UMUX) Scale. *Journal of Usability Studies*, *11*(3), 89-109.

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, *31*(8), 484-495. https://doi.org/10.1080/10447318.2015.1064648

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49-59. https://doi.org/10.1016/0005-7916(94)90063-9

Brooke, J. (1996). SUS- A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4-7. https://doi.org/10.1201/9781498710411-35

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29-40.

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64-73. https://doi.org/10.1177/002224377901600110

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates. https://doi.org/10.4324/9780203771587

Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*(3), 289-304. https://doi.org/10.1086/269376

Dolbier, C. L., Webster, J. A., McCalister, K. T., Mallon, M. W., & Steinhardt, M. A. (2005). Reliability and validity of a single-item measure of job satisfaction. *American Journal of Health Promotion*, *19*(3), 194-198. https://doi.org/10.4278/0890-1171-19.3.194

Fedele, G., Fedriga, M., Zanuso, S., Mastrangelo, S., & Di Nocera, F. (2017). *Can User Experience affect buying intention? A case study on the evaluation of exercise equipment.* Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference, Prague, Czech Republic.

Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies*, *1*(4), 185-188.

Finstad, K. (2010). The Usability Metric for User Experience. *Interacting with Computers*, *22*(5), 323-327. https://doi.org/10.1016/j.intcom.2010.04.004

Fisher, R. A. (1925). *Statistical Methods for Research Workers* (1 ed.). Oliver and Boyd.

Gräve, E., Bell, R., & Buchner, A. (2024). Verbal and pictorial single-item scales are as good as their 10-item counterparts for measuring perceived usability. *Ergonomics*, 1-15. https://doi.org/10.1080/00140139.2024.2371061

Gräve, E., & Buchner, A. (2024). Is less sometimes more? An experimental comparison of four measures of perceived usability. *Human Factors*, *0*(0). https://doi.org/10.1177/00187208241237862

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185. https://doi.org/10.1007/BF02289447

ISO-9241-110. (2008). Ergonomics of human-system interaction - Part 110: Dialogue principles (ISO 9241-110:2006); German version EN ISO 9241-110:2006. *Deutsches Institut für Normung, Berlin*.

ISO-9241-210. (2010). Ergonomics of human–system interaction–Part 210: Human-centred design for interactive systems. In: ISO Geneva, Switzerland.

ISO-9241/10. (1995). *Ergonomic requirements for office work with visual display terminals (VDTs) – Part 10: Dialogue principles*.

ISO-9241/11. (1998). *Ergonomic requirements for work with visual display terminals (VDTs)-Part 11: Guidance on usability*.

Kortum, P., Acemyan, C. Z., & Oswald, F. L. (2021). Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). *Human Factors*, *63*(6), 987-998. https://doi.org/10.1177/0018720819881556

Kortum, P., & Oswald, F. L. (2018). The impact of personality on the subjective assessment of usability. *International Journal of Human-Computer Interaction*, *34*(2), 177-186. https://doi.org/10.1080/10447318.2017.1336317

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology*, *8*(1), 65-77. https://doi.org/10.1111/j.1744-6570.1955.tb01189.x

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction*, *34*(12), 1148-1156. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. R. (2019). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human–Computer Interaction*, *35*(15), 1404-1419. https://doi.org/10.1080/10447318.2018.1533152

Lewis, J. R., & Sauro, J. (2017). Revisiting the factor structure of the System Usability Scale. *Journal of Usability Studies*, *12*(4), 183-192.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). *Umux-Lite - When there's no time for the SUS.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496-505. https://doi.org/10.1080/10447318.2015.1064654

Marcus, A. (2005). User interface design's return on investment: Examples and statistics. In R. G. Bias & D. J. Mayhew (Eds.), *Cost-Justifying Usability* (pp. 17-39). Elsevier. https://doi.org/10.1016/B978-012095811-5/50002-X

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77-86. https://doi.org/10.1348/096317902167658

Nielsen, J. (1994). *Usability engineering*. Elsevier. https://doi.org/10.1016/C2009-0-21512-1

Nunnally, J. C. (1978). *Psychometric Theory* (Second ed.). McGraw-Hills.

Oshagbemi, T. (1999). Overall job satisfaction: How good are single versus multiple-item measures? *Journal of Managerial Psychology*, *14*(5), 388-403. https://doi.org/10.1108/02683949910277148

Pomeroy, I. M., Clark, C. R., & Philp, I. (2001). The effectiveness of very short scales for depression screening in elderly medical patients. *International Journal of Geriatric Psychiatry*, *16*(3), 321-326. https://doi.org/10.1002/gps.344

Prümper, J. (1993). Software-evaluation based upon ISO 9241 Part 10. In Grechenig & Tscheligi (Eds.), *Human Computer Interaction. VCHCI 1993.* (Vol. 733, pp. 255-265). Springer, Berlin ,Heidelberg. https://doi.org/10.1007/3-540-57312-7_74

Prümper, J. (1997). The ISONORM 9241/10 usability questionnaire: Results on reliability and validity. In R. Liskowsky, B. M. Velichkovsky, & W. Wünschmann (Eds.), *Software-Ergonomie '97: Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung* (pp. 253–262). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-86782-7_21

Prümper, J. (1999). Test IT: ISONORM 9241/10. In H. J. Bullinger & J. Ziegler (Eds.), *Human-Computer Interaction – Communication, Cooperation, and Application Design* (pp. 1028-1032). Lawrence Erlbaum Associates.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305-335. https://doi.org/10.1016/S0167-8116(02)00097-6

Sauro, J. (2018). Can you use a single item to predict SUS scores? *Measuring U*. https://measuringu.com/single-item-sus/

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Proceedings of the SIGCHI conference on human factors in computing systems, Vancouver, BC, Canada*. https://doi.org/10.1145/1978942.1979266

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245-251. https://doi.org/10.1037/0033-2909.87.2.245

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: how good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247-252. https://doi.org/10.1037/0021-9010.82.2.247

Einzelarbeiten

Einzelarbeit 1

Die Einzelarbeit enthält die Experimente 1a und 1b.
Gräve, E., & Buchner, A. (2024). Is less sometimes more? An experimental comparison of four measures of perceived usability. *Human Factors*. https://doi.org/10.1177/00187208241237862

Research Article

Is Less Sometimes More? An Experimental Comparison of Four Measures of Perceived Usability

Human Factors
2024, Vol. 0(0) 1–17

DOI: 10.1177/00187208241237862
journals.sagepub.com/home/hfs
Sage

Elisa Gräve and Axel Buchner

Abstract

Objective: In usability studies, the subjective component of usability, perceived usability, is often of interest besides the objective usability components, efficiency and effectiveness. Perceived usability is typically investigated using questionnaires. Our goal was to assess experimentally which of four perceived-usability questionnaires differing in length best reflects the difference in perceived usability between systems.

Background: Conventional measurement wisdom strongly favors multi-item questionnaires, as measures based on more items supposedly yield better results. However, this assumption is controversial. Single-item questionnaires also have distinct advantages and it has been shown repeatedly that single-item measures can be viable alternatives to multi-item measures.

Method: $N = 1089$ (Experiment 1) and $N = 1095$ (Experiment 2) participants rated the perceived usability of a good or a poor web-based mobile phone contract system using the 35-item ISONORM 9241/10 (Experiment 1 only), the 10-item System Usability Scale (SUS), the 4-item Usability Metric for User Experience (UMUX), and the single-item Adjective Rating Scale.

Results: The Adjective Rating Scale represented the perceived-usability difference between both systems at least as good as, or significantly better than, the multi-item questionnaires (significantly better than the UMUX and the ISONORM 9241/10 in Experiment 1, significantly better than the SUS in Experiment 2).

Conclusion: The single-item Adjective Rating Scale is a viable alternative to multi-item perceived-usability questionnaires.

Application: Extremely short instruments can be recommended to measure perceived usability, at least for simple user interfaces that can be considered concrete-singular in the sense that raters understand which entity is being rated and what is being rated is reasonably homogenous.

Keywords

ISONORM, system usability scale, usability metric for user experience, adjective rating scale, standardized questionnaires

Introduction

When evaluating user interfaces of systems such as computer programs or online stores, objective metrics such as error rates and the time taken to complete a task are important, but so are the

Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Received: September 14, 2022; accepted: February 17, 2024

Corresponding Author:
Elisa Gräve, Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstraße 1, Düsseldorf 40225, Germany; e-mail: elisa.graeve@hhu.de

impressions resulting from using the system, collectively referred to as user experience (ISO-9241-210, 2010). Among other things, the user experience can influence buying intentions (Fedele et al., 2017) and how likely users recommend a system to others (Brooke, 2013). A fundamental component of the experience of users with a system is captured by the construct of perceived usability (Lewis, 2018). In addition to the objective aspects, effectiveness and efficiency, the subjective component of usability forms a large part of the construct of usability (ISO-9241/11, 1998). A cost-efficient approach to measuring perceived usability is to use standardized questionnaires. An array of pertinent questionnaires is available which vary in length and complexity. The goal of the research presented here is to test experimentally which of several such questionnaires best reflects the difference between a good and a poor system in terms of perceived usability.

One rather long questionnaire is the 35-item ISONORM 9241/10 (Prümper, 1993). It is based on the corresponding ISO norm (ISO-9241/10, 1995) and thus has a solid theoretical foundation. Seven sets of five items were designed to measure the extent to which a system fulfills the seven design requirements formulated in the ISO 9241/10 norm (Table A1 in the appendix). The ISONORM 9241/10 has been shown to be a valid instrument to measure the subjective component of usability (Prümper, 1997, 1999).

The 10-item System Usability Scale (SUS, see Brooke, 1996, 2013) was developed as a quick instrument for assessing perceived usability (Table A1 shows the version of the SUS with exclusively positive wording, see below). Over the years, the SUS has become the gold standard among perceived-usability questionnaires, most likely in large part because Bangor et al. (2008) showed, based on data collected in over a decade, that the SUS is a valid and very robust instrument, has the ability to measure changes in usability during an iterative design cycle and is suitable for a wide range of systems and types of technology. In the standard form of the SUS, items have alternating positive and negative wording. Sauro and Lewis (2011) changed the negative wording to positive wording to avoid errors such as misinterpretations of negative items. As the standard and the positively worded SUS yield essentially the same results, the positively worded SUS is recommended especially for nonsupervised studies (Kortum et al., 2021; Sauro & Lewis, 2011).

Even shorter than the SUS, the Usability Metric for User Experience (UMUX; Table A1) consists of only four items. These are based on the ISO-9241/11 (1998) definition of usability (Finstad, 2010) and have alternating positive and negative wording. Research indicates acceptable levels of concurrent validity of the UMUX, which is why it is often considered a shorter alternative to the SUS (Berkman & Karahoca, 2016; Finstad, 2010; Lewis, 2013).

Bangor et al. (2009) added an item to the SUS with an adjective-anchored Likert scale, the Adjective Rating Scale (Table A1). They administered the SUS together with this single item to 964 participants. The single-item measure correlated highly with the SUS ($r = .822$). A priori it may seem inadequate to use a single-item questionnaire to arrive at a measure of perceived usability but given the high correlation with the SUS, the Adjective Rating Scale might be an economical alternative to multi-item perceived-usability questionnaires.

However, conventional measurement wisdom strongly favors the use of multi-item questionnaires because, for instance, measures based on many items are often considered to be more reliable than measures based on few items. Then again, long questionnaires can also have disadvantages. First, measuring perceived usability with a complex multi-item questionnaire may require participants to rate irrelevant or unknown aspects of a system. We will get back to this issue in the General Discussion. Second, long survey instruments are time-consuming, require a considerable investment of motivation and effort and may even overload the respondents (Wanous et al., 1997), leading to a decline in data quality and quantity (Dillman et al., 1993), particularly in online surveys (Baumgartner et al., 2019b; Evans & Mathur, 2005).

From this perspective, single-item measures have several advantages. They require little time

and are easy to administer (Pomeroy et al., 2001) which may increase the willingness to complete a questionnaire (Wanous et al., 1997). What is more, it has been shown repeatedly that single-item measures can have acceptable psychometric properties that are often equivalent to those of their multi-item counterparts. For instance, Himmels et al. (2021) have shown that an ad-hoc single-item measure can be as good as, or better than, the SUS and the UMUX when assessing user reactions to an automated driving system, and Christophersen and Konradt (2011) have found strong indications for the reliability, validity and sensitivity of a single-item measure of online store usability. Research in other disciplines has yielded similar findings. For instance, single-item measures yield valid measures of stress symptoms (Eddy et al., 2019; Wong et al., 2021), overall job satisfaction (Dolbier et al., 2005; Nagy, 2002) and life satisfaction (Cheung & Lucas, 2014; Jovanović & Lazić, 2020). Single-item measures have also been validated in fields as diverse as health care (de Boer et al., 2004), sports (Bruton et al., 2016; Kwon & Trail, 2005), marketing (Ang & Eisend, 2018; Bergkvist & Rossiter, 2007, 2009) and organizational psychology (Matthews et al., 2022). Thus, valid single-item measures are not uncommon (for principled arguments in favor of single-item measures, see Allen et al., 2022; Fuchs & Diamantopoulos, 2009).

Rossiter (2002) has argued that, in general, single-item measures are appropriate whenever the to-be-measured construct and its attribute can be considered to concern entities that are concrete and singular in the sense that (a) all raters understand which entity is being rated and (b) what is being rated is reasonably homogenous. With respect to perceived usability, it seems possible that systems with relatively simple user interfaces such as automated driving systems or web shops may be considered concrete-singular (Bangor et al., 2009).

Given these considerations, the question arises whether questionnaires with more or fewer items are optimal for measuring the construct of perceived usability. To answer this question, the ISONORM 9241/10, the SUS, the UMUX and the Adjective Rating Scale were compared experimentally. As prior experience with a system can influence the evaluation of the system's perceived usability (Berkman & Karahoca, 2016; Borsci et al., 2015), we designed two systems that were novel for all participants. Specifically, we designed two simulated web-based mobile phone contract systems that differed in usability, resulting in what we term here a good-usability system and a poor-usability system. Participants worked with one of the systems and then completed the ISONORM 9241/10 (only in Experiment 1), the positively worded SUS, the UMUX and the Adjective Rating Scale. The aim of the study was to assess which of these four instruments best reflects the difference in perceived usability between the good-usability system and the poor-usability system.

Experiment 1

Method

Statistical power considerations. We used data published by Finstad (2010, p. 326) to guide our statistical power considerations. However, for the population correlations $\rho_{systems,SUS}$ and $\rho_{systems,UMUX}$ (the correlations between the variable coding the type of system [good-usability system vs. poor-usability system] and the SUS and the UMUX, respectively) and $\rho_{SUS,UMUX}$ (the correlation between the SUS and the UMUX), we chose values below those of the corresponding sample correlations reported by Finstad (2010) for two reasons. First, sample correlations tend to overestimate population correlations. Second, we used a commercial research panel and therefore expected our data to be noisier than the data obtained by Finstad with Intel® employees. On this basis, an a priori power analysis using G*Power (Faul et al., 2007, 2009) suggested that given $\alpha = .05$, a power of $1 - \beta = .95$ and assumed population correlations of $\rho_{systems,SUS} = .70$, $\rho_{systems,UMUX} = .65$ and $\rho_{SUS,UMUX} = .80$ we needed data from $N = 1009$ participants for the statistical comparisons of the sample correlations $r_{systems,SUS}$ and $r_{systems,\ UMUX}$.

We stopped collecting data at the end of the day at which the number of valid data sets surpassed this goal.

Participants and Design. In total, 1920 participants living in Germany were recruited using the research panel of Respondi AG, Cologne, Germany, and gave informed consent, but 825 did not complete the study. Of these 345 dropped out while working with the good-usability system and 398 while working with the poor-usability system. Six data sets had to be excluded because of multiple participation. The final sample included 1089 participants (546 male, 541 female, two diverse) with a mean age of 46 years (SD = 15). Participants received a small monetary compensation for their participation. They were randomly assigned to the good-usability system (n = 569) or the poor-usability system (n = 520). All participants filled out four usability questionnaires.

Materials

All experiments reported here were conducted online, implemented using SoSci Survey (Leiner, 2022) and made available via https://www.soscisurvey.de. Participation was possible only with a laptop or desktop computer. We designed two simulated web-based mobile phone contract systems with two goals in mind. First, the systems should require user input. Second, it should be possible to complete the interactions with the system within the time tolerable in an online study. The good-usability system was implemented so that it complied with a common standard of the dialog principles of human–system interaction (ISO-9241-110, 2008). The poor-usability system violated many of these principles (Figure 1). We selected those principles that were easy to comply with or violate in an online study. The differences between the systems concerned, for instance, the text-background polarity (Buchner & Baumgartner, 2007; Piepenbrock et al., 2013, 2014), the spatial arrangement of the dialog elements and the comprehensibility of error messages and instructions (screenshots of all pages of the systems' user interface of both Experiment 1 (translated into English) and 2 are available online, see the Data Availability Statement).

To measure perceived usability, the ISONORM 9241/10 (Prümper, 1993), the positively worded SUS (Sauro & Lewis, 2011), the UMUX (Finstad, 2010) and the Adjective Rating Scale (Bangor et al., 2009) were implemented in SoSci Survey. The ISONORM 9241/10 was available in German. The SUS, the UMUX and the Adjective Rating Scale were translated into German by us.

Procedure

At the beginning of each experiment, written informed consent was obtained from all participants. They knew that they could withdraw their consent at any time by closing their browser window. The experiments were carried out in accordance with the Declaration of Helsinki. If participants decided to continue, they entered their age, gender and education level. Next, they worked with one of the two simulated web-based mobile phone contract systems. Participants knew that they interacted with a mock system. Their task was to perform the actions necessary to purchase a mobile phone contract using this system. They selected a particular plan, could select optional items (e.g., they could choose a second SIM card) and entered data of a fictitious person (what was entered was not recorded) and so on. This took about 5 minutes. Afterward, participants evaluated their experience with the system using the ISONORM 9241/10, the SUS, the UMUX and the Adjective Rating Scale which were presented in a different random order to every participant.

Results

The average usability scores calculated for the four perceived-usability questionnaires are shown in Table 1. To assess the difference in usability between the good-usability system and the poor-usability system, correlations were computed between the variable coding the type of system on the one side and the usability scores of each of the four

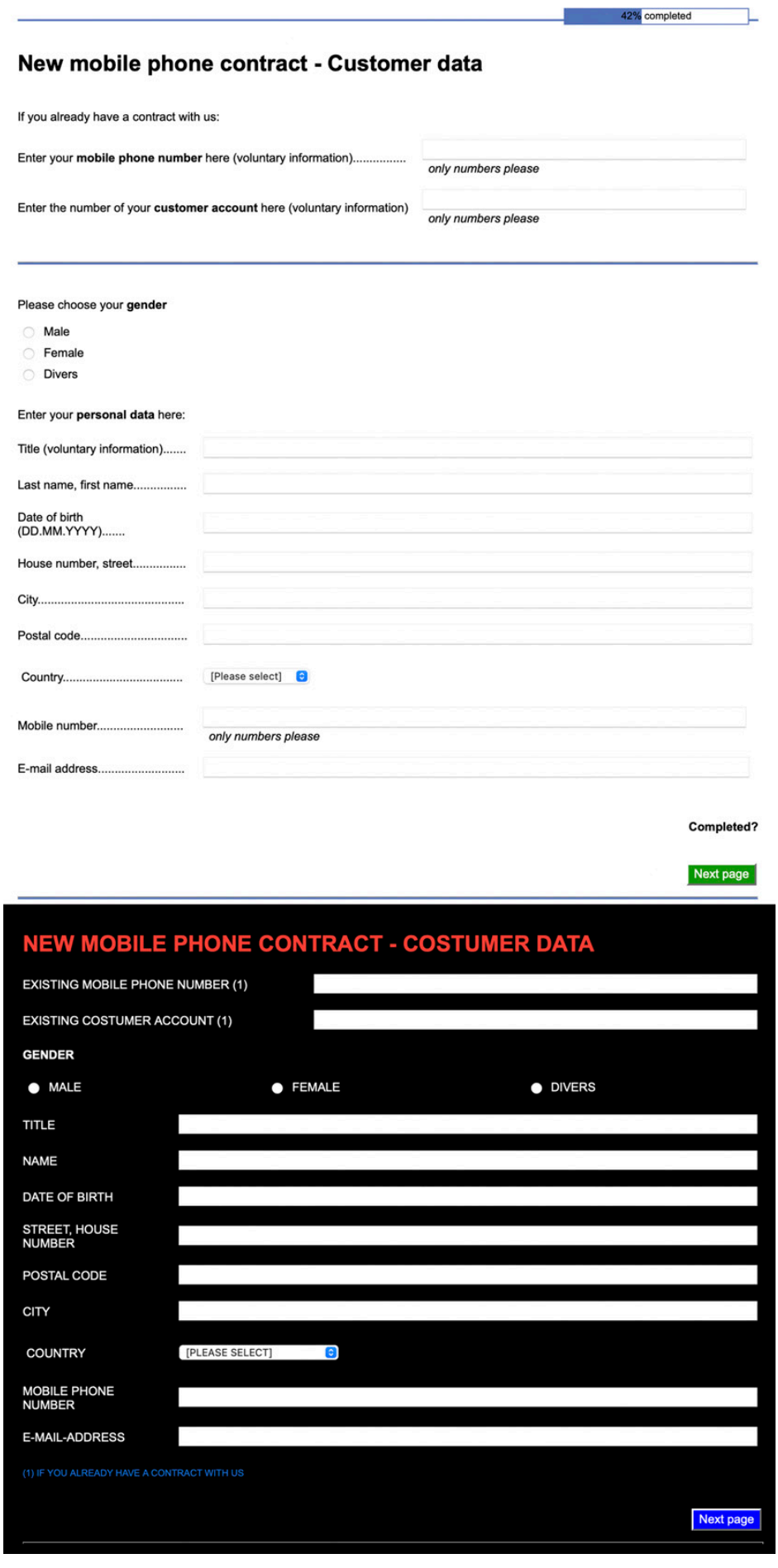


Figure 1. Illustration of displays for the good-usability system (top) and the poor-usability system (bottom) used in Experiment 1 (the original interface language was German).

perceived-usability questionnaires on the other. These correlations, also reported in Table 1, serve two purposes.

First, these correlations represent sample effect size measures, that is, measures of the size of the difference between the good-usability system and the poor-usability system as reflected in a particular perceived-usability measure (for convenience, the size of the perceived-usability difference between the two systems is also shown in terms of the effect size measure *d* in Table 1). All correlations are significantly different from zero, that is, all four measures represent the difference in usability between the good-usability system and the poor-usability system.

Second, these correlations can be compared to each other statistically using the Steiger (1980) test for differences between dependent and overlapping correlations. The correlations are dependent because all participants completed all questionnaires, and the correlations are overlapping because one variable—the variable coding the type of system—is part of both of the to-be-compared correlations. We used the Steiger (1980) test as implemented in *cocor* (Diedenhofen & Musch, 2015). The test requires as input the sample effect size *r* for each of two to-be-compared measures (Table 1) and the correlations between the to-be-compared measures (Table 2). The latter correlations are relatively high which seems largely compatible with the assumption that these measures reflect the same construct, albeit the correlations are clearly lower than the $r = .96$ between the SUS and the UMUX reported by Finstad (2010). Another observation is that the correlations among the perceived-usability measures within each of the two systems (Tables B1 and B2 in the Appendix) are in the same order of magnitude as the correlations reported in Table 2 and are not zero or negative so that a Simpson's Paradox (Simpson, 1951) can be ruled out. To avoid alpha error accumulation within the family of six two-sided tests needed for all pairwise comparisons of the four perceived-usability measures, we used the Bonferroni-Holm procedure (Holm, 1979). The results are shown in Table 3.

First, the Adjective Rating Scale reflected the difference in perceived usability between the good-usability system and the poor-usability system significantly better than the UMUX. Second, the SUS was better than the UMUX at representing this difference. Third, the Adjective Rating Scale was better than the ISONORM. Fourth, although the SUS would be classified as

Table 1. Means (and Standard Deviations in Parentheses) of the Perceived-Usability Measures for the Four Perceived-Usability Questionnaires (Possible Range of Values in Brackets) Separately for the Good-Usability System and the Poor-Usability System, the Effect Size *r* Representing the Size of the Difference Between the Good-Usability System and the Poor-Usability System, the 95% Confidence Intervals of *r*, the *p* Values Associated With the Test of the H0 That $r = 0$ (i.e., That the Two Systems do Not Differ in Perceived Usability) and, for Convenience, the Effect Size Measure *d* Representing the Size of the Difference Between the Good-Usability System and the Poor-Usability System (Experiment 1).

	ISONORM 9241/10 [−3 to 3]	SUS [0 to 100]	UMUX [0 to 100]	Adjective Rating Scale [0 to 100]
Good-usability system	1.54 (1.06)	81.49 (16.07)	82.41 (18.74)	71.76 (16.05)
Poor-usability system	0.90 (1.29)	69.52 (20.65)	72.20 (22.14)	59.39 (20.25)
Effect size *r*	.261	.309	.242	.322
95% confidence interval	.205 to .316	.254 to .362	.186 to .297	.268 to .374
p Value of the test that $r = 0$	<.001	<.001	<.001	<.001
Effect size *d*	0.541	0.650	0.499	0.680

Note. For the ISONORM, responses on a scale from −3 to +3 are coded as such, summed up and the total is divided by 35. For the SUS, responses on a scale from 1 to 5 are coded as values between 0 and 4, summed up and the total is multiplied by 2.5. For the UMUX, responses on a scale from 1 to 7 are coded as values between 0 and 6 for odd items and as values between 6 and 0 for even items, all values are summed up and the total is divided by 24 and multiplied by 100. For the Adjective Rating Scale, responses on a scale from 1 to 7 are coded as values between 0 and 6, divided by 6 and multiplied by 100. For all scales, higher values indicate better usability.

Table 2. Correlations Among the Perceived-Usability Measures Across the Two Systems and Their 95% Confidence Intervals in Parentheses (Experiment 1).

	ISONORM 9241/10	SUS	UMUX
SUS	.728 (.699–.755)	1	
UMUX	.624 (.587–.659)	.706 (.675–.735)	1
Adjective Rating Scale	.719 (.689–.747)	.714 (.683–.742)	.616 (.577–.651)

Table 3. Results of the Tests of Differences Between the Perceived-Usability Measures in How Well They Reflect the Difference in Perceived Usability Between the Good-Usability System and the Poor-Usability System (Steiger's [1980] Test of Differences Between Dependent and Overlapping Correlations; Experiment 1).

	z	p	α_{crit}
Adjective Rating Scale versus UMUX	3.164	.002	.0083
SUS versus UMUX	3.015	.003	.01
Adjective Rating Scale versus ISONORM	2.823	.005	.0125
SUS versus ISONORM	2.252	.024	.0167
ISONORM versus UMUX	0.751	.453	.025
Adjective Rating Scale versus SUS	0.602	.547	.05

significantly better than the ISONORM 9241/10 without control for alpha error accumulation, this difference must be regarded as not statistically significant within the present family of tests. Finally, the ISONORM 9241/10 was not significantly better than the UMUX and the Adjective Rating Scale was not better than the SUS at representing the perceived-usability difference.

Discussion

One half of the results of Experiment 1 is rather unsurprising: The SUS is significantly better than the UMUX at representing the differences between a good-usability system and a poor-usability system. Similar results have already been presented by Finstad (2010) when he introduced the UMUX. His "System 2" was more usable than his "System 1" and the SUS ($r = .89$) was better than the UMUX ($r = .86$) at representing the difference in usability between these systems. Finstad did not evaluate this difference statistically. When this is done, the results show that the SUS is significantly better than the UMUX at representing the difference between the two systems used in that study, $z = 5.36$, $p < 0.001$.

The other half of the results is more surprising. The 35-item ISONORM 9241/10 is no better than the 4-item UMUX at representing the difference between systems. This is the opposite of what would be expected based on conventional measurement wisdom. The ISONORM 9241/10 thus seems to be the least efficient choice for measuring perceived usability, delivering no better discriminability than shorter questionnaires at a higher cost (with more items). We thus decided not to consider the ISONORM 9241/10 any further.

The biggest surprise is that the single-item Adjective Rating Scale reflected the usability differences between the two systems used in Experiment 1 so well. At a descriptive level, the Adjective Rating Scale reflected the difference between the good-usability system and the poor-usability system better than the other measures with more items. When evaluated statistically, the Adjective Rating Scale was significantly better than the UMUX and the ISONORM 9241/10 and as good as the SUS. This result confirms that, at least under certain circumstances, single-item measures may be better at capturing perceived usability than previously thought.

Table 4. Means (and Standard Deviations in Parentheses) of the Perceived-Usability Measures for the Three Perceived-Usability Questionnaires (Possible Range of Values in Brackets) Separately for the Good-Usability System and the Poor-Usability System, the Effect Size *r* Representing the Size of the Difference Between the Good-Usability System and the Poor-Usability System, the 95% Confidence Intervals of *r*, the *p* Values Associated With the Test of the H0 That $r = 0$ (i.e., That the Two Systems do Not Differ in Perceived Usability) and, for Convenience, the Effect Size Measure *d* Representing the Size of the Difference Between the Good-Usability System and the Poor-Usability System (Experiment 2).

	SUS [0 to 100]	UMUX [0 to 100]	Adjective Rating Scale [0 to 100]
Good-usability system	75.86 (19.62)	75.45 (21.75)	69.09 (18.51)
Poor-usability system	50.74 (25.83)	47.33 (27.29)	42.30 (25.70)
Effect size *r*	.481	.494	.517
95% confidence interval	.434 to .525	.448 to .537	.472 to .559
p Value of the test that $r = 0$	<.001	<.001	<.001
Effect size *d*	1.097	1.136	1.208

The goal of Experiment 2 was to test whether the results of Experiment 1 can be replicated conceptually. The following changes were made. First, a sample of participants living in Britain responded to the original English language versions of the SUS, the UMUX and the Adjective Rating Scale. Second, we modified the mobile phone contract systems to differ even more in usability than they did in Experiment 1. Applying the effect size conventions suggested by (Cohen, 1988) to the correlations reported in Table 1 one may conclude that the difference in perceived usability between the good-usability system and the poor-usability system was about "medium." Our goal for Experiment 2 was to create a "large" difference between the two systems. To anticipate, the measures we took to achieve this goal (see below) were successful (Table 4).

Experiment 2

Method

Statistical power considerations, participants, and design. The statistical power considerations were the same as those of Experiment 1. A total of 2644 participants living in Britain were recruited using the research panel of Respondi AG, Cologne, Germany, and gave informed consent, but 1549 did not complete the study. Of these 554 dropped out while working with a system with good usability and 833 while working with a system with poor usability. The final sample included 1095 participants (693 male, 395 female, 7 diverse) with a mean age of 54 years (SD = 15). Participants received a small monetary compensation for their participation. They were randomly assigned to the good-usability system (n = 688) and the poor-usability system (n = 407). All participants filled out three perceived-usability questionnaires.

Materials, procedure, and design. Materials, procedure, and design were the same as those of Experiment 1 with the following exceptions. The user interfaces of the simulated web-based mobile phone contract systems were translated into English. Additional measures were taken to increase the difference in usability between the two systems. For instance, the poor-usability system was made even less usable by reducing the font-background contrast and by mixing right-justified and left-justified text. Finally, the ISONORM 9241/10 was dropped due to its undesirable properties (see above).

Results

The average usability scores calculated for the three perceived-usability questionnaires are shown in Table 4, as are the correlations between the variable coding the type of system on the one side and the usability scores of each of the three perceived-usability questionnaires on the other. All correlations are significantly different from zero, that is, all three measures represent the differences in usability between the good-usability system and the poor-usability system.

Table 5. Correlations Among the Perceived-Usability Measures Across the Two Systems and Their 95% Confidence Intervals in Parentheses (Experiment 2).

	SUS	UMUX
UMUX	.819 (.798–.838)	1
Adjective Rating Scale	.845 (.827–.861)	.801 (.779–.821)

To statistically compare two perceived-usability measures we again used the Steiger (1980) test for differences between dependent and overlapping correlations as implemented in *cocor* (Diedenhofen & Musch, 2015). The test requires as input the sample effect size *r* for each of two to-be-compared measures (Table 4) and the correlations between the measures (Table 5). The latter correlations are again relatively high which seems largely compatible with the assumption that these measures capture the same construct. The correlations among the perceived-usability measures within each of the two systems (Tables B3 and B4 in the Appendix) are in the same order of magnitude as the correlations reported in Table 5 and are not zero or negative, so that a Simpson's Paradox (Simpson, 1951) can be ruled out. To avoid alpha error accumulation within the family of three two-sided tests needed for all pairwise comparisons of the three perceived-usability measures we used the Bonferroni–Holm procedure (Holm, 1979). The results are presented in Table 6. First, the Adjective Rating Scale reflected the difference in perceived usability between the good-usability system and the poor-usability system significantly better than the SUS. Second, the Adjective Rating Scale was not significantly better than the UMUX. Third, the SUS was not significantly better than the UMUX.

Discussion

As in Experiment 1, the Adjective Rating Scale reflected the difference between the good-usability system and the poor-usability system descriptively better than the other measures. However, the difference was statistically significant only for the comparison between the Adjective Rating Scale and the SUS, not for the comparison between the Adjective Rating Scale and the UMUX. The fact that the Adjective Rating Scale reflected the perceived-usability differences so well twice in the present series of experiments strengthens the assumption that the Adjective Rating Scale may be appropriate, if not optimal, for perceived-usability assessments at least in some circumstances, as will be discussed below.

General Discussion

The most prominent aspect of the data obtained in the present experiments is that the most economical measure of perceived usability, the single-item Adjective Rating Scale (Bangor et al., 2009), reflected the differences in usability between two systems at least as good as, and in some cases even significantly better than, established multi-item measures. At a descriptive level, the Adjective Rating Scale was better than all other measures considered here in both Experiment 1 and Experiment 2. In Experiment 1 the Adjective Rating Scale was significantly better than the UMUX and the ISONORM 9241/10 and was as good as the SUS. In Experiment 2, the Adjective Rating Scale was significantly better than the SUS and as good as the UMUX. Considering this and the fact that measurements of perceived usability cannot be any simpler than with a single item, we conclude that the Adjective Rating Scale seems to be the most efficient instrument for assessing perceived usability.

As mentioned in the introduction, successful single-item scales are not uncommon and have been found to yield valid measurements in many areas. Rossiter (2002) has argued that a single-item measure may be appropriate whenever the to-be-

Table 6. Results of the Tests of Differences Between the Perceived-Usability Measures in How Well They Reflect the Difference in Perceived Usability Between the Good-Usability System and the Poor-Usability System (Steiger's [1980] Test of Differences Between Dependent and Overlapping Correlations; Experiment 2).

	z	p	α_{crit}
Adjective Rating Scale versus SUS	2.50	.013	.0167
Adjective Rating Scale versus UMUX	1.421	.155	.025
SUS versus UMUX	0.83	.407	.05

measured construct and its attribute can be considered to concern entities that are concrete and singular in the sense that (a) all raters understand which entity is being rated and (b) what is being rated is reasonably homogenous.

In the case considered here it seems plausible that condition (a) is fulfilled (Bangor et al., 2009). The next question is whether perceived usability is a reasonably homogeneous construct. As an indicator of homogeneity, we analyzed whether the SUS and the UMUX can be considered unidimensional measures by conducting an exploratory factor analysis, assuming that a single-factor solution to the factor extraction problem would indicate unidimensionality of the measures and, hence, the measured construct. A parallel analysis (Horn, 1965) using principle-component extraction and retaining all factors corresponding to eigenvalues greater than the 95th percentile of the reference eigenvalues (Auerswald & Moshagen, 2019) consistently revealed one significant dimension for both the SUS (Experiment 1: eigenvalue = 6.37, explaining 64% of the variance; Experiment 2: eigenvalue = 7.19, explaining 72% of the variance) and the UMUX (Experiment 1: eigenvalue = 2.49, explaining 62% of the variance; Experiment 2: eigenvalue = 2.84, explaining 71% of the variance). We conclude that both the SUS and the UMUX can be treated as essentially unidimensional which is consistent with recent conclusions by others (Berkman & Karahoca, 2016; Kortum et al., 2021; Lewis, 2019; Lewis & Sauro, 2017). We consider this to be an indicator of the homogeneity of the construct of perceived usability in the sense of Rossiter (2002).

Given the success of the single-item Adjective Rating Scale it seems appropriate to speculate about why this is the case. A possible reason is that a single-item text such as that of the Adjective Rating Scale ("Overall, I would rate the user-friendliness of this product as…") is relatively general and open to interpretation. This may enable users to particularly flexibly focus on the aspect of the system which is the most salient determinant of perceived usability in a given context (e.g., the high clarity of the good-usability systems and the particularly bad legibility of the all-capitals negative polarity text in the poor-usability systems used here), thereby maximizing the difference between systems in terms of perceived-usability ratings. Whether this really is the case is a question that affords an empirical answer in an independent study.

Furthermore, as mentioned in the Introduction it seems possible that with longer perceived-usability questionnaires participants are required to respond to items that cover irrelevant or unknown aspects of the system in question. For instance, after having interacted with an online store interface, participants may not have a basis for answering items aiming at the possibilities to automate frequently recurring operations (Table A1, ISONORM 9241/10, Item 3), the controlling of information presented on the screen (Table A1, ISONORM 9241/10, Item 24), the integration of functions in the system (Table A1, SUS, Item 5) or the correcting of things with a system (Table A1, UMUX, Item 4). It is not clear what happens in such cases. Some participants may respond to such items with ratings they already gave to more relevant items. Alternatively, some participants could opt for a neutral response option. However, what really determines the responses in such cases is ultimately an empirical question. At this stage it seems plausible that such responses do not increase, and may in fact weaken, the ability of a perceived-usability questionnaire to discriminate between systems with good and bad usability.

Probably the biggest limitation of single-item measures of perceived usability is that they only inform about the level of perceived usability of a system or the difference in usability between systems while the reasons for this evaluation remain obscure. Therefore, such single-item

measures are of limited utility during formative evaluation. Here, multi-item questionnaires consisting of several dimensions should be more useful than a single item because multi-item questionnaires should provide more information about where usability weaknesses might be located (Baumgartner et al., 2019a). However, the Adjective Rating Scale may turn out to be useful as a quick indicator of whether newly added features or changes made to a system in response to user feedback represent progress or regression in perceived usability, similar to what Bangor et al. (2008) have demonstrated for the SUS (see Table 12 in Bangor et al., 2008).

The next question is how one would predict precisely which perceived-usability measure—single-item or multiple-item—best reflects the difference between good and poor usability for a specific system. At present it seems that a precise answer to this question is not possible and that, in the end, the best answer will have to be given at an empirical level. However, as a heuristic for practitioners it seems reasonable to expect that single-item measures of perceived usability may be appropriate whenever the to-be-evaluated system is so simple that its interface can be considered concrete-singular, as defined by Rossiter (2002). This seems to apply, for example, to interfaces with only a main navigation and no help system (Bangor et al., 2009) such as the simulated web-based mobile phone contract systems used here. However, how well this heuristic works is an empirical question that needs to be answered in future studies.

An obvious limitation of the present experiments is that every participant filled out all perceived-usability questionnaires, albeit this occurred in a different random order for each participant. Such an approach is efficient, but the downside could be the risk of carry-over effects in the sense that the completion of earlier perceived-usability questionnaires may influence the way in which later questionnaires are completed. However, the chances of finding statistically significant carry-over effects within the present experimental design are close to zero even if carry-over effects existed. There were 24 and 6 different groups defined by the random sequences in which participants completed the perceived-usability questionnaires in Experiments 1 and 2, respectively. This would result in 276 and 15 statistical tests of pairwise group differences in Experiments 1 and 2, respectively. The statistical power of each test of a group difference would be minimal given the small sample sizes of the groups and the extremely small levels of alpha needed to control for alpha error accumulation within these large families of statistical tests. To avoid carry-over effects, one person should complete only one perceived-usability questionnaire. However, in this case the sample size requirements can easily become very large which means that such investigations will realistically be limited to a comparison of two perceived-usability questionnaires.

Another limitation of the present study is that the well-evaluated, positively worded SUS was used, whereas the UMUX contained two positive and two negative items. The reason for this is that no evaluated version for the UMUX with only positive items was available. Therefore, the question remains as to whether a version of the UMUX with exclusively positive items would perform better in nonsupervised perceived-usability studies than the version of the UMUX used here.

Conclusion

Less may be more in measuring perceived usability. The single-item Adjective Rating Scale (Bangor et al., 2009) is the shortest of the perceived-usability measures considered here which at the same time reflected the difference in perceived usability between systems at least as well as the UMUX (Finstad, 2010), the SUS (Brooke, 1996) and the ISONORM 9241/10 (Prümper, 1993). In Experiment 1 the Adjective Rating Scale was significantly better at representing the difference in usability between a good-usability system and a poor-usability system than the UMUX and the ISONORM 9241/10, in Experiment 2 it was significantly better than SUS and at a descriptive level it was better than all other measures of perceived usability in both experiments.

It thus seems that the Adjective Rating Scale can be recommended as an efficient and valid alternative to the total scores of subjective-usability questionnaires (i.e., when more detailed feedback is not needed or helpful for the design of the interface) and when the to-be-rated interfaces can be considered concrete-singular (Rossiter, 2002).

Appendix A

Table A1. All Items of the Four Usability Questionnaires used: the ISONORM 9241/10 (Translated Into English by the Authors of This Article), the Version of the SUS With Positively Worded Items, the UMUX, and the Adjective Rating Scale.

	Items
ISONORM 9241/10	*Suitability for the task*: Does the system support the completion of your work tasks without burdening you unnecessarily?
	The system…
	1. is/is not complicated to use.
	2. does not offer/offers all functions to complete the tasks at hand efficiently.
	3. offers poor/good possibilities to automate frequently recurring processing operations.
	4. does not require/requires superfluous input.
	5. is poorly/well tailored to the requirements of the work.
	Self-descriptiveness: Does the system give you enough explanations and is it sufficiently understandable?
	The system…
	6. provides a poor/good overview of its range of functions.
	7. uses terms, designations, abbreviations or symbols in masks and menus that are difficult/easy to understand.
	8. provides insufficient/sufficient information about which entries are permitted or necessary.
	9. does not give/give situation-specific explanations on request that are of concrete help.
	10. does not provide/provides situation-specific and helpful explanations when requested.
	Conformity with user expectations: Does the system meet your expectations and habits through a consistent and understandable design?
	The system…
	11. makes orientation difficult/easy through an inconsistent/a consistent design.
	12. leaves/does not leave you in the dark about whether an entry was successful or not.
	13. provides insufficient/sufficient information about what it is doing.
	14. responds with poorly-predictable/well-predictable processing times.
	15. cannot/can be used consistently according to a uniform principle.
	Suitability for learning: Is the system designed in such a way that you could easily familiarize yourself with it and does it also offer support when you want to learn new functions?
	The system…
	16. requires a lot of/little time to learn.
	17. does not encourage/encourages you to try new functions.
	18. requires/does not require to remember many details.
	19. is designed so that what is learned is difficult/easy to remember.
	20. is difficult/easy to learn without outside help or a manual.
	Controllability: Can you influence the way you work with the system?
	The system…
	21. offers no/the possibility to interrupt work at any point and continue there later without losses.
	22. forces an/no unnecessarily rigid adherence to processing steps.
	23. does not make/makes it possible to switch easily between individual menus or masks.
	24. is designed in such a way that the user cannot/can influence how and what information is presented on the screen.
	25. enforces/does not force unnecessary interruptions of the work.
	Error tolerance: Does the system offer you the possibility to achieve the intended work result with little or no correction effort despite incorrect entries?
	The system …
	26. is designed in such a way that small errors can/cannot have serious consequences.
	27. informs too late/immediately about faulty entries.
	28. provides error messages that are difficult/easy to understand.
	29. overall requires a high/low level of correction effort when errors occur.
	30. does not give/gives concrete advice on how to correct errors.
	Suitability for individualization: Can you as a user adapt the system to your individual needs and requirements without much effort?
	The system…
	31. is difficult/easy for me to extend when new tasks arise.
	32. can be adapted badly/well by me to my personal, individual way of working.
	33. is not/is equally suitable for beginners and experts because I can adapt it to my level of knowledge with difficulty/ease.
	34. can hardly/easily be adapted by me—within the scope of its possibilities—to different tasks.
	35. is designed in such a way that I can hardly/easily adapt the screen display to my individual needs.

(continued)

Table A1. (continued)

		Items
SUS	1.	I think that I would like to use this system frequently.
	2.	I found the system to be simple.
	3.	I thought the system was easy to use.
	4.	I think I could use the system without the support of a technical person.
	5.	I found the various functions in the system were well integrated.
	6.	I thought there was a lot of consistency in the system.
	7.	I would imagine that most people would learn to use this system very quickly.
	8.	I found the system very intuitive.
	9.	I felt very confident using the system.
	10.	I could use the system without having to learn anything new.
UMUX	1.	This system's capabilities meet my requirements.
	2.	Using this system is a frustrating experience.
	3.	This system is easy to use.
	4.	I have to spend too much time correcting things with this system.
	1.	Overall, I would rate the user-friendliness of this product as:
Adjective Rating Scale		• Worst imaginable • Awful • Poor • Ok • Good • Excellent • Best Imaginable

Appendix B

Table B1. Correlations Among the Usability Measures Within the *Good-Usability System* and Their 95% Confidence Intervals in Parentheses (Experiment 1).

	ISONORM 9241/10	SUS	UMUX
SUS	.662 (.614–.706)	1	
UMUX	.524 (.461–.581)	.635 (.583–.681)	1
Adjective Rating Scale	.637 (.585–.683)	.672 (.624–.715)	.516 (.454–.574)

Table B2. Correlations Among the Usability Measures Within the *Poor-Usability System* and Their 95% Confidence Intervals in Parentheses (Experiment 1).

	ISONORM 9241/10	SUS	UMUX
SUS	.736 (.694–.773)	1	
UMUX	.656 (.605–.703)	.721 (.677–.760)	1
Adjective Rating Scale	.737 (.695–.774)	.689 (.641–.732)	.636 (.582–.685)

Table B3. Correlations Among the Usability Measures Within the *Good-Usability System* and Their 95% Confidence Intervals in Parentheses (Experiment 2).

	SUS	UMUX
UMUX	.713 (.674–.748)	
Adjective Rating Scale	.762 (.728–.791)	.683 (.641–.721)

Table B4. Correlations Among the Usability Measures Within the *Poor-Usability System* and Their 95% Confidence Intervals in Parentheses (Experiment 2).

	SUS	UMUX
UMUX	.814 (.778–.844)	
Adjective rating scale	.825 (.791–.853)	.784 (.743−.819)

Key Points

- Four common perceived-usability questionnaires were experimentally validated and compared: the 35-item ISONORM 9241/10 (only in Experiment 1), the 10-item System Usability Scale (SUS), the 4-item Usability Metric for User Experience (UMUX) and the single-item Adjective Rating Scale.
- The Adjective Rating Scale was as good as, and sometimes significantly better than, all other measures of perceived usability considered here at representing the difference in usability between a good-usability system and a poor-usability system.
- The Adjective Rating Scale can be recommended as a particularly efficient and valid measure of perceived usability, at least for interfaces that are considered concrete-singular (Rossiter, 2002) such as the simulated web-based mobile phone contract systems used here (Bangor et al., 2009).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Elisa Gräve https://orcid.org/0000-0003-4653-7498

Data Availability Statement

The data of both experiments as well as screenshots of the simulated web-based mobile phone contract systems used here (translated into English for the systems used in Experiment 1) are available at the project page of the Open Science Framework under https://osf.io/k8dbe/

References

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment*, *38*(1), 1–5. https://doi.org/10.1027/1015-5759/a000699

Ang, L. & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, *58*(2), 218–227. https://doi.org/10.2501/JAR-2017-001

Auerswald, M. & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. https://doi.org/10.1037/met0000200

Bangor, A., Kortum, P., & Miller, J. (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

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Bangor_May2009.pdf

Baumgartner, J., Frei, N., Kleinke, M., Sauer, J., & Sonderegger, A. (2019a). Pictorial System Usability Scale (P-SUS): Developing an instrument for measuring perceived usability. In CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland, UK. ACM. https://doi.org/10.1145/3290605.3300299

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019b). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies*, *122*, 78–89. https://doi.org/10.1016/j.ijhcs.2018.08.008

Bergkvist, L. & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175–184. https://doi.org/10.1509/jmkr.44.2.175

Bergkvist, L. & Rossiter, J. R. (2009). Tailor-made single-item measures of doubly concrete constructs. *International Journal of Advertising*, *28*(4), 607–621. https://doi.org/10.2501/S0265048709200783

Berkman, M. I. & Karahoca, D. (2016). Re-assessing the Usability Metric for User Experience (UMUX) Scale. *Journal of Usability Studies*, *11*(3), 89–109. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Berkman_May2016.pdf

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, *31*(8), 484–495. https://doi.org/10.1080/10447318.2015.1064648

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7. https://doi.org/10.1201/9781498710411-35

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies*, *8*(2), 29–40. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Brooke_February_2013.pdf

Bruton, A. M., Mellalieu, S. D., & Shearer, D. A. (2016). Validation of a single-item stem for collective efficacy measurement in sports teams. *International Journal of Sport and Exercise Psychology*, *14*(4), 383–401. https://doi.org/10.1080/1612197X.2015.1054853

Buchner, A. & Baumgartner, N. (2007). Text–background polarity affects performance irrespective of ambient illumination and colour contrast. *Ergonomics*, *50*(7), 1036–1063. https://doi.org/10.1080/00140130701306413

Cheung, F. & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality of Life Research*, *23*(10), 2809–2818. https://doi.org/10.1007/s11136-014-0726-4

Christophersen, T. & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies*, *69*(4), 269–280. https://doi.org/10.1016/j.ijhcs.2010.10.005

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates. https://doi.org/10.4324/9780203771587

de Boer, A. G. E. M., van Lanschot, J. J. B., Stalmeier, P. F. M., van Sandick, J. W., Hulscher, J. B. F., de Haes, J. C. J. M., & Sprangers, M. A. G. (2004). Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research*, *13*(2), 311–320. https://doi.org/10.1023/B:QURE.0000018499.64574.1f

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One*, *10*(3), Article e0121945. https://doi.org/10.1371/journal.pone.0121945

Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*(3), 289–304. https://doi.org/10.1086/269376

Dolbier, C. L., Webster, J. A., McCalister, K. T., Mallon, M. W., & Steinhardt, M. A. (2005). Reliability and validity of a single-item measure of job satisfaction. *American Journal of Health Promotion: AJHP*, *19*(3), 194–198. https://doi.org/10.4278/0890-1171-19.3.194

Eddy, C. L., Herman, K. C., & Reinke, W. M. (2019). Single-item teacher stress and coping measures: Concurrent and predictive validity and sensitivity to change. *Journal of School Psychology*, *76*, 17–32. https://doi.org/10.1016/j.jsp.2019.05.001

Evans, J. R. & Mathur, A. (2005). The value of online surveys. *Internet Research*, *15*(2), 195–219. https://doi.org/10.1108/10662240510590360

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fedele, G., Fedriga, M., Zanuso, S., Mastrangelo, S., & Di Nocera, F. (2017). Can user experience affect buying intention? A case study on the evaluation of exercise equipment. In Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference. Prague, Czech Republic.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004

Fuchs, C. & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Betriebswirtschaft*, *69*(2), 195–210. https://temme.wiwi.uni-wuppertal.de/fileadmin/_migrated/content_uploads/fuchs_diamantopoulos_2009.pdf

Himmels, C., Omozik, K., Jarosch, O., & Buchner, A. (2021). Measuring user experience in automated driving: Developing a single-item measure. 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. Leeds, United Kingdom. ACM. https://doi.org/10.1145/3409118.3475135

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

ISO-9241/10. (1995). *Ergonomic requirements for office work with visual display terminals (VDTs) – Part 10: Dialogue principles*. ISO.

ISO-9241/11. (1998). *Ergonomic requirements for work with visual display terminals (VDTs)-Part 11: Guidance on usability*. ISO.

ISO-9241-110. (2008). *Ergonomics of human-system interaction - Part 110: Dialogue principles (ISO 9241-110:2006); German version EN ISO 9241-110: 2006*: Deutsches Institut für Normung.

ISO-9241-210. (2010). *Ergonomics of human–system interaction–Part 210: Human-centred design for interactive systems*. ISO Geneva.

Jovanović, V. & Lazić, M. (2020). Is longer always better? A comparison of the validity of single-item versus multiple-item measures of life satisfaction. *Applied Research in Quality of Life*, *15*(3), 675–692. https://doi.org/10.1007/s11482-018-9680-6

Kortum, P., Acemyan, C. Z., & Oswald, F. L. (2021). Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). *Human Factors*, *63*(6), 987–998. https://doi.org/10.1177/0018720819881556

Kwon, H. & Trail, G. (2005). The feasibility of single-item measures in sport loyalty research. *Sport Management Review*, *8*(1), 69–89. https://doi.org/10.1016/s1441-3523(05)70033-4

Leiner, D. J. (2022). *SoSci survey* (Version 3.3.04). [Computer software]. Available at. https://www.soscisurvey.de

Lewis, J. R. (2013). Critical review of 'The usability metric for user experience'. *Interacting with Computers*, *25*(4), 320–324. https://doi.org/10.1093/iwc/iwt013

Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*, *34*(12), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. R. (2019). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human-Computer Interaction*, *35*(15), 1404–1419. https://doi.org/10.1080/10447318.2018.1533152

Lewis, J. R. & Sauro, J. (2017). Revisiting the factor structure of the System Usability Scale. *Journal of Usability Studies*, *12*(4), 183–192. https://uxpajournal.org/de/revisit-factor-structure-system-usability-scale/.

Matthews, R. A., Pineault, L., & Hong, Y.-H. (2022). Normalizing the use of single-item measures: Validation of the single-item compendium for organizational psychology. *Journal of Business and Psychology*, *37*(4), 639–673. https://doi.org/10.1007/s10869-022-09813-3

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77–86. https://doi.org/10.1348/096317902167658

Piepenbrock, C., Mayr, S., & Buchner, A. (2014). Positive display polarity is particularly advantageous for small character sizes: Implications for display design. *Human Factors*, *56*(5), 942–951. https://doi.org/10.1177/0018720813515509

Piepenbrock, C., Mayr, S., Mund, I., & Buchner, A. (2013). Positive display polarity is advantageous for both younger and older adults. *Ergonomics*, *56*(7), 1116–1124. https://doi.org/10.1080/00140139.2013.790485

Pomeroy, I. M., Clark, C. R., & Philp, I. (2001). The effectiveness of very short scales for depression screening in elderly medical patients. *International Journal of Geriatric Psychiatry*, *16*(3), 321–326. https://doi.org/10.1002/gps.344

Prümper, J. (1993). Software-evaluation based upon ISO 9241 Part 10. In Grechenig, & Tscheligi (Eds.), *Human computer interaction. VCHCI 1993* (Vol. *733*, pp. 255–265). Springer. https://doi.org/10.1007/3-540-57312-7_74

Prümper, J. (1997). The ISONORM 9241/10 usability questionnaire: Results on reliability and validity. In R. Liskowsky, B. M. Velichkovsky, & W. Wünschmann (Eds.), *Software-Ergonomie '97: Usability engineering: integration von mensch-computer-interaktion und software-entwicklung* (pp. 253–262). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-86782-7_21

Prümper, J. (1999). Test IT: ISONORM 9241/10. In H. J. Bullinger, & J. Ziegler (Eds.), *Human-computer interaction – communication, cooperation, and application design* (pp. 1028–1032). Lawrence Erlbaum Associates.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305–335. https://doi.org/10.1016/S0167-8116(02)00097-6

Sauro, J. & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, BC, Canada. https://doi.org/10.1145/1978942.1979266

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B*, *13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*,

87(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Wong, S. M. Y., Lam, B. Y. H., Wong, C. S. M., Lee, H. P. Y., Wong, G. H. Y., Lui, S. S. Y., Chan, K. T., Wong, M. T. H., Chan, S. K. W., Chang, W. C., Lee, E. H. M., Suen, Y. N., Hui, C. L. M., & Chen, E. Y. H. (2021). Measuring subjective stress among young people in Hong Kong: Validation and predictive utility of the single-item subjective level of stress (SLS-1) in epidemiological and longitudinal community samples. *Epidemiology and Psychiatric Sciences*, *30*, Article e61. https://doi.org/10.1017/S2045796021000445

Author Biographies

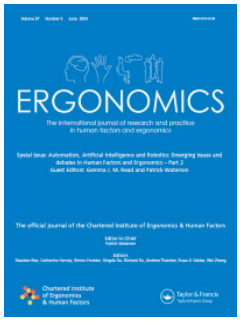
Elisa Gräve is a PhD student in the Department of Experimental Psychology at Heinrich Heine University Düsseldorf in Düsseldorf, Germany.

Axel Buchner is a full professor in the Department of Experimental Psychology at Heinrich Heine University Düsseldorf in Düsseldorf, Germany. He received his PhD in psychology from Bonn University in 1992.

Einzelarbeit 2

Die Einzelarbeit enthält die Experimente 2a und 2b.

Gräve, E., Bell, R., & Buchner, A. (2024). Verbal and pictorial single-item scales are as good as their 10-item counterparts for measuring perceived usability. *Ergonomics.* https://doi.org/10.1080/00140139.2024.2371061



Ergonomics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/terg20

Verbal and pictorial single-item scales are as good as their 10-item counterparts for measuring perceived usability

Elisa Gräve, Raoul Bell & Axel Buchner

To cite this article: Elisa Gräve, Raoul Bell & Axel Buchner (28 Jun 2024): Verbal and pictorial single-item scales are as good as their 10-item counterparts for measuring perceived usability, Ergonomics, DOI: 10.1080/00140139.2024.2371061

To link to this article: https://doi.org/10.1080/00140139.2024.2371061

ERGONOMICS
https://doi.org/10.1080/00140139.2024.2371061

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Verbal and pictorial single-item scales are as good as their 10-item counterparts for measuring perceived usability

Elisa Gräve, Raoul Bell and Axel Buchner

Department of Experimental Psychology, Heinrich Heine University, Düsseldorf, Nordrhein-Westfalen, Germany

ABSTRACT

Single-item scales of perceived usability are attractive due to their efficiency and non-verbal scales are attractive because they enable collecting data from individuals irrespective of their language proficiency. We tested experimentally whether single-item verbal and pictorial scales can compete with their 10-item counterparts at reflecting the difference in usability between well-designed and poorly designed systems. $N = 1079$ (Experiment 1) and $N = 1092$ (Experiment 2) participants worked with two systems whose usability was experimentally manipulated. Perceived usability was assessed using the 10-item System Usability Scale, the single-item Adjective Rating Scale, the 10-item Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale. The single-item scales reflect the difference in usability as good as their 10-item counterparts. The pictorial scales are nearly as valid as their verbal counterparts. The single-item Adjective Rating Scale and the Pictorial Single-Item Usability Scale are thus efficient and valid alternatives to their 10-item counterparts.

Practitioner summary: Verbal and pictorial single-item perceived-usability scales are viable alternatives to their 10-item counterparts. Specifically, the single-item Adjective Rating Scale is as good as the 10-item System Usability Scale and the Pictorial Single-Item Usability Scale is as good as the Pictorial System Usability Scale at reflecting differences in usability between systems.

ARTICLE HISTORY
Received 9 January 2024
Accepted 6 June 2024

KEYWORDS
System Usability Scale; Adjective Rating Scale; Pictorial System Usability Scale; Pictorial Single-Item Usability Scale; perceived usability

Introduction

Short questionnaires are attractive because they are more efficient than longer questionnaires in terms of data collection and analysis and because they increase the chances of high response rates by avoiding survey fatigue. The most efficient questionnaires are those that contain only a single item for measuring a construct, ideally without sacrificing validity. In fact, the validity of single-item scales can be surprisingly good and single-item scales have been successfully validated for constructs in a wide variety of domains (Table 1). Here we focus on the measurement of perceived usability. In short, perceived usability is the subjective ease with which people interact with a system. Perceived usability is crucial when developing and improving interactive systems such as websites, software applications and technical devices. Initial experimental evidence (Gräve and Buchner 2024) suggests that a single-item scale for measuring perceived usability, the Adjective Rating Scale (Bangor, Kortum, and Miller 2009), reflects the difference in perceived usability between well-designed and poorly designed systems as well as, if not better than, multi-item perceived-usability questionnaires: the 4-item Usability Metric for User Experience (Finstad 2010), the 10-item System Usability Scale (Brooke 1996, 2013) and the 35-item ISONORM 9241/10 (Prümper 1993). The present experiments replicate and extend these findings by demonstrating that verbal and pictorial single-item scales of perceived usability reflect the difference in usability between well-designed and poorly designed systems at least as well as their 10-item counterparts. From a practical viewpoint, this is encouraging news in that it shows that perceived usability can be measured very efficiently without sacrificing validity.

Perceived usability is a central aspect of the user experience (Lewis 2018), a term coined to reflect the quality of a user's past or anticipated future use of a product, system or service (ISO-9241-210210, 2010). Perceived usability affects, for instance, how likely people intend to buy a system or recommend it to

CONTACT Elisa Gräve elisa.graeve@hhu.de Department of Experimental Psychology Heinrich Heine University, Düsseldorf, Nordrhein-Westfalen, Germany

Table 1. Examples of constructs for which valid single-item scales exist.

Construct	Exemplary publication
Academic anxiety and academic self-concept	Gogol et al. (2014)
Alcohol misuse	Seale et al. (2006)
Attitudes in advertising research	Ang and Eisend (2018); Bergkvist and Rossiter (2007, 2009)
Body size perception	Truby and Paxton (2002)
Chronotype	de Bruijn et al. (2022)
Cognitive change	Fardell et al. (2022)
Collective efficiency in sports	Bruton, Mellalieu, and Shearer (2016)
Community connectedness	Mashek, Cannaday, and Tangney (2007)
Depressed mood	McKenzie and Marks (1999)
Experience with a simulated automated driving system	Himmels et al. (2021)
Fear of cancer recurrence	Smith et al. (2023)
Focus of attention at work	Gardner et al. (1998)
General health	DeSalvo et al. (2006)
Global self-esteem	Robins, Hendin, and Trzesniewski (2001)
Growth mindset	Rammstedt et al. (in press)
Human energy	Weigelt et al. (2022)
Life satisfaction	Cheung and Lucas (2014); Jovanović and Lazić (2020)
Motivation to stop smoking	Pashutina et al. (2021)
Nausea intensity	Baxter et al. (2011)
Online-store usability	Christophersen and Konradt (2011)
Organizational identification	Shamir and Kark (2004)
Organizational justice	Jordan and Turner (2008)
Overall job satisfaction	Dolbier et al. (2005); Nagy (2002); Oshagbemi (1999); Wanous, Reichers, and Hudy (1997)
Pain intensity	Bieri et al. (1990)
Patient mood	Lorish and Maisiak (1986)
Processing fluency	Graf, Mayer, and Landwehr (2017)
Quality of life	de Boer et al. (2004)
Rehabilitation adherence	Kuo et al. (2023)
Self-compassion	Zhang et al. (2022)
Sleepiness	Maldonado, Bentley, and Mitchell (2004)
Sport loyalty	Kwon and Trail (2005)
Stress and coping	Eddy, Herman, and Reinke (2019)
Stress symptoms	Elo, Leppänen, and Jahkola (2003)
Subjective stress	Wong et al. (2021)

others (Brooke 2013; Fedele et al. 2017). Being able to efficiently measure perceived usability is thus important when developing and improving interactive user interfaces. Measuring perceived usability is also relatively easy when appropriate questionnaires are available.

A well-established questionnaire for assessing perceived usability is the System Usability Scale (Brooke 1996, 2013). The original System Usability Scale (Table A1) comprises five positively and five negatively formulated items, each to be rated on a 5-point Likert scale (e.g. 'I found the system very intuitive'). Based on data collected in over a decade, the System Usability Scale is considered a 'highly robust and versatile tool for usability professionals' (Bangor, Kortum, and Miller 2008, p. 574), suitable for evaluating a wide range of interactive systems. In fact, the System Usability Scale can be considered the gold standard for measuring perceived usability. A variant of the System Usability Scale with only positively formulated items (Table A2) has been developed to avoid misinterpretations of the negatively formulated items and confusion that may result from mixing positively and negatively formulated items (Sauro and Lewis 2011). Both versions of the System Usability Scale have yielded parallel results in the past but due to its straightforward design, the positively worded System Usability Scale is recommended for situations in which minimal instruction is provided (Kortum, Acemyan, and Oswald 2021).

The Adjective Rating Scale (Table A3) can be considered a single-item alternative to the System Usability Scale. Originally introduced by Bangor, Kortum, and Miller (2009), the Adjective Rating Scale has been demonstrated experimentally to reflect the difference between a well-designed system and a poorly designed system as well as, or even better than, several multi-item scales (Gräve and Buchner 2024). Thus, the Adjective Rating Scale has been demonstrated to come at no cost in terms of validity while providing the pragmatic advantages of a single-item scale.

However, a limitation of this previous experimental evaluation is that the scales considered are verbal scales. When verbal scales are used, persons with limited proficiency in the language in which the questionnaire is written may be excluded from evaluating a system or may misinterpret the wording of items and thus give unintended responses (Bradley and Lang 1994; Kunin 1955). To allow for a more inclusive assessment of perceived usability, Baumgartner, Frei, et al. (2019) introduced the Pictorial System Usability Scale (Table A4), a pictorial version of the original System Usability Scale (Table A1). Each pictorial item of this scale is based on one of the items of the original System Usability Scale and is rated on a bipolar 7-point Likert scale with pictorial representations of their positive and negative endpoints. These pictorial representations consist of an avatar and several other elements designed to support the communication of the meaning of the items. A validation study with a relatively small sample ($N = 60$) yielded encouraging results: the correlation between the pictorial and the verbal System Usability Scale total scores was $r = .865$ (Baumgartner, Frei, et al. 2019).

As with the original System Usability Scale, there is a single-item alternative to the 10-item Pictorial System Usability Scale (Baumgartner, Sonderegger, et al. 2019). The Pictorial Single-Item Usability Scale (Table A5) consists of a single bipolar 9-point Likert scale ranging from −4 to 4. Certain points on the scale are labelled by pictorial elements designed to support the communication of the meaning of the scale: Moving from the negative to the positive endpoint of the scale, a thumb-down gesture turns into a thumb-up gesture, a

face with an angry expression turns into a face with a happy expression and the background colour of the scale elements changes from saturated red to white to saturated green. In addition, an icon representing the to-be-evaluated system is presented. A validation study with a relatively small sample ($N = 38$) yielded encouraging results: the correlation between the Pictorial Single-Item Usability Scale score and the verbal System Usability Scale total score was $r = .696$ (Baumgartner, Sonderegger, et al. 2019).

In the experiments reported below we compared these four perceived-usability scales, the 10-item System Usability Scale, the single-item Adjective Rating Scale, the 10-item Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale. More precisely, we tested experimentally how well these scales reflect the difference in perceived usability between two systems designed to differ in usability. Based on the facts that the validity of single-item scales can be surprisingly good, that single-item scales have already been successfully validated for constructs in a variety of domains and that Gräve and Buchner (2024) have presented initial experimental evidence showing that perceived usability can be measured very efficiently with a single item without compromising validity, the central hypothesis to be tested here was that single-item measures of perceived usability reflect the difference in usability between a well-designed and poorly designed system as well as their 10-item counterparts. This should hold true for both verbal and pictorial scales.

Participants first interacted with one of two novel simulated web-based mobile phone contract purchasing systems, a well-designed system and a poorly designed system. Participants then judged the perceived usability of the system they had interacted with using a 10-item and a single-item perceived usability scale, both of which were either verbal or pictorial scales. Our first goal was to test whether it was possible to replicate the finding that the single-item Adjective Rating Scale reflects the difference in usability between the well-designed system and the poorly designed system at least as well as the original 10-item System Usability Scale (Gräve and Buchner 2024). Our second goal was to test whether a parallel conclusion can be reached for the pictorial alternatives to these verbal scales. Specifically, we tested whether the Pictorial Single-Item Usability Scale reflects the difference in usability between the well-designed system and the poorly designed system as well as the 10-item Pictorial System Usability Scale.

Apart from that, we also tested whether the 10-item Pictorial System Usability Scale reflects the difference in perceived usability between the well-designed system and the poorly designed system at least as well as the 10-item System Usability Scale. Finally, we also tested whether the Pictorial Single-Item Usability Scale reflects the difference in perceived usability between the well-designed system and the poorly designed system at least as well as the single-item Adjective Rating Scale.

Experiment 1

Method

Participants and design

UK residents were recruited with the help of the research panel of Bilendi (https://www.bilendi.de/). Participants received a small monetary compensation for their participation. Before analysing the results, the data sets of participants who did not complete the experiment, were less than 18 years old (and thus could not legally consent to the use of their data in Germany) or had participated twice were excluded from data analysis. Of the participants who had given informed consent, 415 respondents dropped out while working with the well-designed system and 675 dropped out while working with the poorly designed system. The final sample included 1079 participants (589 male, 483 female, 7 diverse) with a mean age of 51 years ($SD = 15$). Of these participants, 1074 reported to speak English fluently and 5 reported not to speak English fluently. The sample reflected a wide range of educational levels: 2 % reported to have no formal degree, 19 % reported to have, as their highest qualification, a General Certificate of Secondary Education or an equivalent qualification, 24 % reported to have an A-Level, an International Baccalaureate or an equivalent qualification and 54 % reported to have a university degree or an equivalent qualification. The participants were randomly assigned to four groups: (1) Participants who evaluated the well-designed system using the 10-item and the single-item verbal scales ($n = 339$), (2) participants who evaluated the poorly designed system using the 10-item and the single-item verbal scales ($n = 204$), (3) participants who evaluated the well-designed system using the 10-item and the single-item pictorial scales ($n = 328$) and (4) participants who evaluated the poorly designed system using the 10-item and the single-item pictorial scales ($n = 208$).

Power analysis

A sensitivity analysis was performed which was focused on the correlations between the total scores of the perceived-usability scales and the variable coding the type of system (well-designed vs. poorly designed; see

the 'System' variable in the datafiles available online as documented in the Data Availability Statement) that are central to the analyses reported below (see Results section). The higher this correlation, the better the perceived-usability scale is at reflecting the difference in usability between the well-designed system and the poorly designed system. To test whether the newly developed pictorial perceived-usability scales differ from the well-established verbal perceived-usability scales at reflecting the difference in usability between the well-designed system and the poorly designed system, the correlation between the total score of a pictorial perceived-usability scale and the variable coding the type of system is to be compared to the correlation between the total score of a verbal perceived-usability scale and the variable coding the type of system using a two-sided Fisher (1925) test for independent correlations. A sensitivity analysis using G*Power (Faul et al. 2009; Faul et al., 2007) showed that given levels of $\alpha = \beta = .05$, $N_{\text{verbal scale}} = 543$ and $N_{\text{pictorial scale}} = 536$, effects of size $q = 0.22$ could be detected. In terms of the effect size conventions introduced by Cohen (1988), this effect size lies between a small ($q = 0.10$) and a medium ($q = 0.30$) effect size.

Ethics statement

The experiment was conducted according to the ethical principles of the Declaration of Helsinki. Participants gave their written informed consent at the start of the experiment. They could withdraw their consent at any time by closing the browser window.

Materials and procedure

System interfaces. The experiment was conducted online using SoSci Survey (Leiner 2022). Participation in the experiment was only possible with a laptop or desktop computer. Participants were informed that they were going to interact with a mock online mobile phone contract purchasing system. They were told to imagine that they wanted to purchase a new mobile phone contract using the system. They could select one of several contracts as well as optional extra services (e.g. a second SIM card) and were asked to enter fictitious personal data to complete the purchase (none of these data were stored).

It was randomly determined whether participants interacted with the well-designed system or the poorly designed system. Figure 1 displays example pages of the well-designed and the poorly designed system (screenshots of all pages of the systems' user interface are available online, see the Data Availability Statement). The well-designed system largely conformed to established human-system interaction dialogue principles (ISO-9241/1010, 1995) whereas the poorly designed system violated many of these principles. As in a previous validation study (Gräve and Buchner 2024), principles were chosen that were easy to conform to or to violate in a web-based interface. Examples include the provision or omission of progress feedback (Cao, Ritz, and Raad 2013; Gronier et al. 2019; Scapin and Bastien 1997), good or poor comprehensibility of instructions and error messages (Lohr 2000; Molich and Nielsen 1990), lowercase versus all-caps text (Tinker 1955), high or low colour contrast (Hall and Hanna 2004; Lin 2003) and positive versus negative text-background polarity (Buchner and Baumgartner 2007; Piepenbrock, Mayr, and Buchner 2014; Piepenbrock et al. 2013).

Perceived-usability measures. After having interacted with the online mobile phone contract purchasing system, participants evaluated their experience with it. Two verbal instruments—the positively worded 10-item System Usability Scale (Sauro and Lewis 2011) (Table A2) and the single-item Adjective Rating Scale (Bangor, Kortum, and Miller 2009) (Table A3)—and two pictorial instruments—the 10-item Pictorial System Usability Scale (Baumgartner, Frei, et al. 2019) (Table A4) and the Pictorial Single-Item Usability Scale (Baumgartner, Sonderegger, et al. 2019) (Table A5)—were used to measure perceived usability. It was randomly determined whether a participant received the verbal perceived-usability scales or the pictorial perceived-usability scales and whether the 10-item perceived-usability scale or the single-item usability scale was presented first. For both the verbal and the pictorial scales participants received a single-sentence instruction asking them to rate their experience with the system they had been working with based on the statement(s) or picture(s) that would be presented next. The System Usability Scale and Pictorial System Usability Scale ratings were aggregated into a total score that was rescaled to span the interval from 0 to 100; the Adjective Rating Scale and the Pictorial Single-Item Usability Scale ratings were rescaled analogously (for details, see the notes of Tables A1 to A5 in the Appendix).

Results

The data of Experiment 1 are available online (see the Data Availability Statement). In Table 2, the average total scores of the positively worded 10-item System Usability Scale, the single-item Adjective Rating Scale, the 10-item Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale are reported for the well-designed system and the poorly designed system. To assess how well the perceived-usability scales

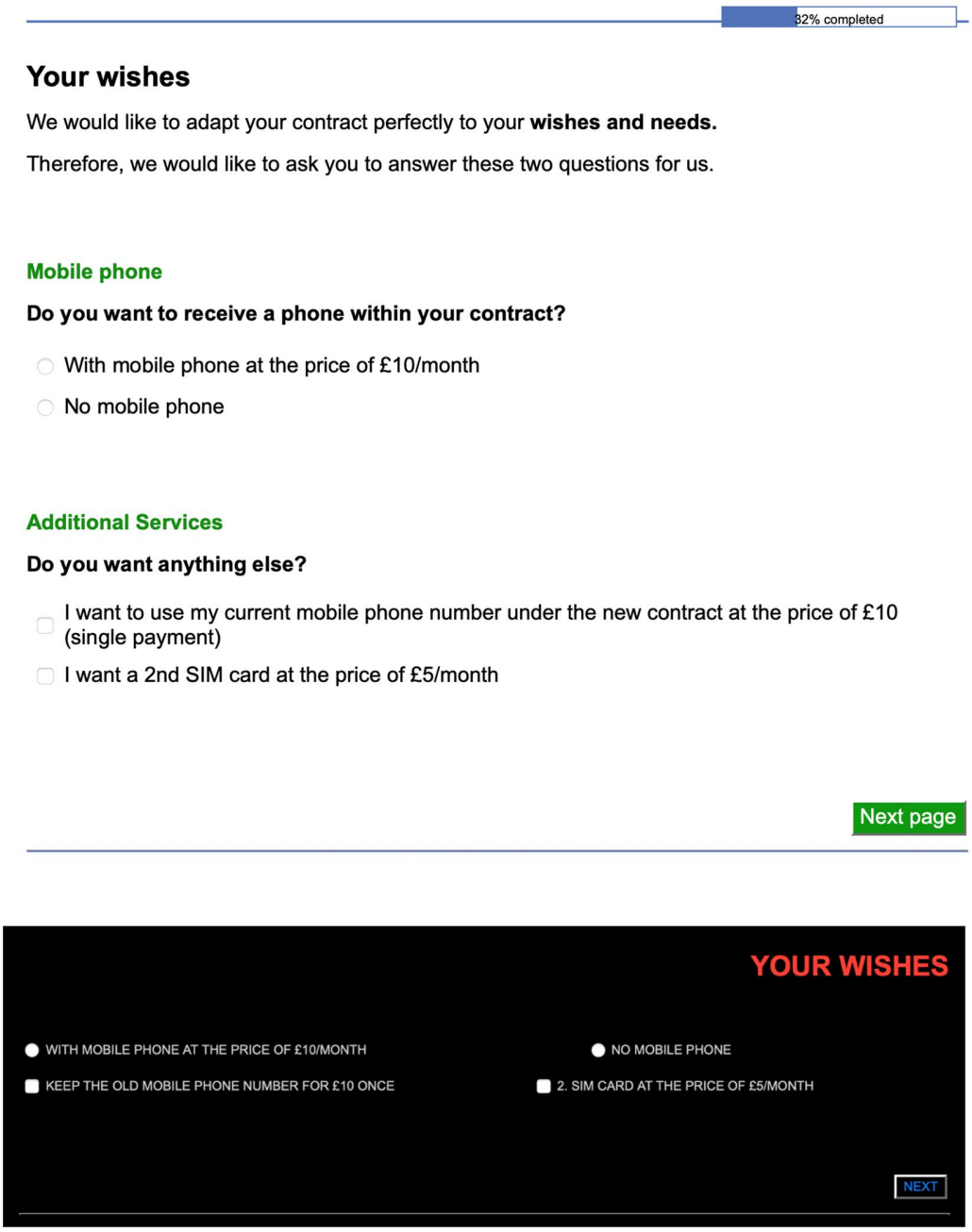


Figure 1. Example pages of the user interfaces for the well-designed system (top) and the poorly designed system (bottom).

reflect the difference in usability between the two systems, correlations were calculated between the total score of each perceived-usability scale and the variable coding the type of system (well-designed system vs. poorly designed system). These correlations, reported in Table 2, represent the size of the difference in perceived usability between the well-designed system and the poorly designed system. The larger the correlation, the better the perceived-usability scale is at reflecting the difference in usability between the two systems. As shown in Table 2, the correlations are significantly different from zero for all perceived-usability scales (all p's $< .001$), leading to the conclusion that all four perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system. Descriptively the usability difference is best reflected by the Adjective Rating Scale, followed by the positively worded 10-item System Usability Scale, the Pictorial Single-Item Usability Scale and the 10-item Pictorial System Usability Scale.

In Table 3, the correlations between the total scores of the 10-item perceived-usability scales and the

Table 2. Statistical indicators of how well the perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system in Experiment 1.

	System Usability Scale	Adjective Rating Scale	Pictorial System Usability Scale	Pictorial Single-Item Usability Scale
Perceived usability of the well-designed system	78.61 (17.65)	72.86 (16.45)	72.47 (18.06)	77.93 (22.66)
Perceived usability of the poorly designed system	60.07 (24.54)	53.02 (25.37)	60.46 (24.40)	61.18 (30.54)
Sample effect size *r*	.402	.429	.272	.300
95 % confidence interval	.329 to .470	.358 to .495	.192 to .349	.221 to .375
p for test of H_0: $r=0$	< .001	< .001	< .001	< .001

Note: Means of the total scores of the positively worded 10-item System Usability Scale, the single-item Adjective Rating Scale, the 10-item Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale are reported for the well-designed system and the poorly designed system. The values in parentheses represent the standard deviations. The sample effect size *r* represents the size of the difference in perceived usability between the well-designed system and the poorly designed system. The 95 % confidence intervals refer to the sample effect size *r*. The *p* values refer to the test of the null hypothesis that $r=0$, that is, that there is no difference in perceived usability between the well-designed system and the poorly designed system.

scores of the single-item perceived-usability scales are reported, separately for the verbal and the pictorial scales. These correlations are needed to perform the Steiger (1980) test for differences between dependent and overlapping correlations which is used to evaluate statistically whether the correlations reported in Table 2 differ significantly from each other (see below). The correlations are dependent because all participants completed either all verbal or all pictorial questionnaires and the correlations are overlapping because one variable—the variable coding the type of system—is part of both correlations being compared. As a side note, the fact that the correlations between the 10-item scales and the single-item scales reported in Table 3 are quite high is consistent with the assumption that both the two verbal scales and the two pictorial scales measure the same construct.

It is already obvious from the relative size of the correlations reported in Table 2 that the single-item Adjective Rating Scale reflects the difference in usability between the well-designed system and the poorly designed system at least as well as the 10-item System Usability Scale: At the descriptive level the single-item Adjective Rating Scale reflects the difference in usability even somewhat better than the 10-item System Usability Scale. A two-sided Steiger (1980) test for differences between dependent and overlapping correlations shows that there is no statistically significant difference between the Adjective Rating Scale and the 10-item verbal System Usability Scale in how well these scales reflect the difference in usability between the well-designed system and the poorly designed system (Table 4). A similar result was obtained for the two pictorial scales. The relative size of the correlations presented in Table 2 shows that, at the descriptive level, the Pictorial Single-Item Usability Scale reflects the difference in usability between the systems even somewhat better than the 10-item Pictorial System Usability Scale. A two-sided Steiger (1980) test for differences between dependent and overlapping correlations shows that there is no statistically significant difference between the Pictorial Single-Item Usability Scale and the Pictorial System Usability Scale in how well these scales reflect the difference in usability between the well-designed system and the poorly designed system (Table 4).

Table 3. Correlations between the 10-item perceived-usability scales and single-item perceived-usability scales in Experiment 1.

	Adjective Rating Scale
System Usability Scale	.768 (.731 – .800)
	Pictorial Single-Item Usability Scale
Pictorial System Usability Scale	.759 (.721 – .793)

Note: The total score of the positively worded 10-item System Usability Scale correlates positively with the score of the single-item Adjective Rating Scale (upper half of the table) and the total score of the Pictorial System Usability Scale correlates positively with the score of the Pictorial Single-Item Usability Scale (lower half of the table) across both systems. The 95 % confidence intervals are reported in parentheses.

Table 4. Results of the statistical test of whether a difference exists in how well the 10-item perceived-usability scales and the single-item perceived-usability scales reflect the usability difference between the well-designed system and the poorly designed system in Experiment 1.

	z	*p*
System Usability Scale vs. Adjective Rating Scale	1.025	.305
Pictorial System Usability Scale vs. Pictorial Single-Item Usability Scale	0.977	.329

Note: The verbal (upper row) and the pictorial (lower row) single-item scales reflect the difference in usability between the well-designed system and the poorly designed system as well as their 10-item counterparts, as shown by the Steiger (1980) test for differences between dependent and overlapping correlations implemented in cocor (Diedenhofen and Musch 2015).

Finally, to test whether there is a difference between the verbal and the pictorial perceived-usability scales in how well they reflect the difference in usability between the well-designed system and the poorly designed system, the positively worded System Usability Scale was compared to the Pictorial System Usability Scale and the Adjective Rating Scale was compared to the Pictorial Single-Item Usability Scale using a two-sided Fisher (1925) test for independent correlations. These correlations are independent as

Table 5. Results of the statistical test of whether a difference exists in how well the verbal perceived-usability scales as opposed to the pictorial perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system in Experiment 1.

	z	*p*
System Usability Scale vs. Pictorial System Usability Scale	2.408	.016
Adjective Rating Scale vs. Pictorial Single-Item Usability Scale	2.443	.015

Note: The 10-item (upper row) and single-item (lower row) verbal perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system better than their pictorial counterparts, as shown by the Fisher (1925) test for independent correlations.

disjunct groups of participants filled out the verbal and the pictorial scales. The results are reported in Table 5. The verbal 10-item System Usability Scale was significantly better than the 10-item Pictorial System Usability Scale and the verbal single-item Adjective Rating Scale was significantly better than the Pictorial Single-Item Usability Scale at reflecting the difference in usability between the two systems.

Discussion

The results of Experiment 1 show that both the verbal and the pictorial single-item scales are not significantly different from their 10-item counterparts at reflecting the difference in usability between the well-designed system and the poorly designed system. At the descriptive level, the single-item scales were even somewhat better at reflecting the difference in usability than their 10-item counterparts. This replicates the findings reported by Gräve and Buchner (2024) for the verbal perceived-usability scales and extends these findings by demonstrating that parallel conclusions hold for the pictorial perceived-usability scales. Further, the verbal perceived-usability scales were significantly better at reflecting the difference in usability between the well-designed system and the poorly designed system than their pictorial counterparts. However, considering that the pictorial scales were still relatively successful at reflecting the difference in usability between the well-designed system and the poorly designed system, we deemed a further experimental comparison between the verbal and pictorial scales necessary before drawing conclusions about the relative usefulness of verbal and pictorial perceived-usability scales.

Experiment 2 was to serve as a conceptual replication of Experiment 1. For that purpose, Experiment 2 was designed to be parallel to Experiment 1 except that the original System Usability Scale (Bangor, Kortum, and Miller 2008; Finstad 2006) was used instead of the positively worded System Usability Scale (Sauro and Lewis 2011) that we had used in Experiment 1. The positively worded version of the System Usability Scale is recommended for non-supervised studies in which it is not possible to help respondents with their questions or to correct their mistakes (Kortum, Acemyan, and Oswald 2021; Sauro and Lewis 2011). This is why the positively worded version of the System Usability Scale seemed ideal for the online environment in which Experiment 1 was implemented. However, it is not clear how much of a problem the alternating of positively and negatively worded items in the original System Usability Scale in an unsupervised online context really is. Thus, even though Experiment 2 was also implemented as an online experiment, we decided to use the original System Usability Scale in which items with odd and even ordinal numbers are positively and negatively worded. In this way the System Usability Scale and the Pictorial System Usability Scale are more closely aligned because, in creating the Pictorial System Usability Scale, the alternating positive and negative wording of the original System Usability Scale had been implemented pictorially. Considering the discussion about the robustness of psychological findings (Open Science Collaboration 2015), we also see a conceptual replication of our results as a necessary step for providing a solid empirical basis for informed conclusions.

Experiment 2

Method

Participants and design

As in Experiment 1, UK residents were recruited with the help of the research panel of Bilendi. Before analysing the results, the data sets of participants who did not complete the experiment, were less than 18 years old or had participated twice were excluded from the analyses. Of the participants who had given informed consent, 414 respondents dropped out while working with the well-designed system and 651 dropped out while working with the poorly designed system. The final sample included 1092 participants (658 male, 429 female, 5 diverse) with a mean age of 55 years ($SD = 15$), none of whom had participated in Experiment 1. Of these participants, 1090 reported to speak English fluently and 2 reported not to speak English fluently. The sample reflected a wide range of educational levels: 5 % reported to have no formal degree, 24 % to have, as their highest qualification, a General Certificate of Secondary Education or an equivalent qualification, 23 % reported to have an A-Level, an International

Baccalaureate or an equivalent qualification and 48 % reported to have a university degree or an equivalent qualification. The participants were randomly assigned to four groups: (1) Participants who evaluated the well-designed system using the 10-item and the single-item verbal scales ($n = 330$), (2) participants who evaluated the poorly designed system using the 10-item and the single-item verbal scales ($n = 220$), (3) participants who evaluated the well-design system using the 10-item and the single-item pictorial scales ($n = 332$), and (4) participants who evaluated the poorly designed system using the 10-item and the single-item pictorial scales ($n = 210$).

Power analysis

A sensitivity analysis was performed that was parallel to that of Experiment 1. To test whether the newly developed pictorial perceived-usability scales differ from the well-established verbal perceived-usability scales at reflecting the difference in usability between the well-designed system and the poorly designed system, the correlation between the total score of a pictorial perceived-usability scale and the variable coding the type of system (see the 'System' variable in the datafiles made available as documented in the Data Availability Statement) is to be compared to the correlation between the total score of a verbal perceived-usability scale and the variable coding the type of system using a two-sided Fisher (1925) test for independent correlations. A sensitivity analysis using G*Power (Faul et al. 2009; Faul et al., 2007) showed that given levels of $\alpha = \beta = .05$, $N_{\text{verbal scale}} = 550$ and $N_{\text{pictorial scale}} = 542$, effects of size $q = 0.22$ could be detected. In terms of the effect size conventions introduced by Cohen (1988), this effect size lies between a small ($q = 0.10$) and a medium ($q = 0.30$) effect size.

Materials and procedure

Materials and procedure were identical to those of Experiment 1 with one exception. Instead of the positively worded System Usability Scale (Sauro and Lewis 2011) used in Experiment 1, the original System Usability Scale (Brooke 1996) was used in the version suggested by Bangor, Kortum, and Miller (2008) and by Finstad (2006) in which Item 8 reads 'I found the system very awkward to use' instead of 'I found the system very cumbersome to use' which had been used originally.

Results

The data of Experiment 2 are available online (see the Data Availability Statement). In Table 6, the average total scores of the original 10-item System Usability Scale, the single-item Adjective Rating Scale, the 10-item Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale are reported for the well-designed system and the poorly designed system. To assess how well the perceived-usability scales reflect the difference in usability between the two systems, correlations were calculated between the total score of each perceived-usability scale and the variable coding the type of system (well-designed system vs. poorly designed system). These correlations, reported in Table 6, represent the size of the difference in perceived usability between the well-designed system and the poorly designed system. The larger the correlation, the better the perceived-usability scale is at reflecting the difference in usability between the two systems. As shown in Table 6, the correlations are significantly different from zero for all perceived-usability scales (all p's $< .001$), leading to the conclusion that all four perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system. Descriptively the usability difference is best reflected in the Adjective Rating Scale, followed by the Pictorial Single-Item Usability Scale, the 10-item Pictorial System Usability Scale and the original 10-item System Usability Scale.

In Table 7, the correlations between the total scores of the 10-item perceived-usability scales and

Table 6. Statistical indicators of how well the perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system in Experiment 2.

	System Usability Scale	Adjective Rating Scale	Pictorial System Usability Scale	Pictorial Single-Item Usability Scale
Perceived usability of the well-designed system	76.28 (17.74)	70.71 (16.30)	74.83 (17.31)	81.74 (21.07)
Perceived usability of the poorly designed system	59.75 (20.91)	50.76 (21.94)	56.27 (23.13)	58.57 (29.29)
Sample effect size r	.392	.463	.417	.418
95 % confidence interval	.318 to .460	.394 to .526	.344 to .484	.346 to .485
p for test of H_0: $r = 0$	< .001	< .001	< .001	< .001

Note: Means of the total scores of the original 10-item System Usability Scale, the single-item Adjective Rating Scale, the 10-item Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale are reported for the well-designed system and the poorly designed system. The values in parentheses represent the standard deviations. The sample effect size r represents the size of the difference in perceived usability between the well-designed system and the poorly designed system. The 95 % confidence intervals refer to the sample effect size r. The p values refer to the test of the null hypothesis that $r = 0$, that is, that there is no difference in perceived usability between the well-designed system and the poorly designed system.

the scores of the single-item perceived-usability scales are reported, separately for the verbal and the pictorial scales. These correlations are needed to perform the Steiger (1980) test for differences between dependent and overlapping correlations which is used to evaluate statistically whether the correlations reported in Table 6 differ significantly from each other (see below). The correlations are dependent because all participants completed either all verbal or all pictorial questionnaires and the correlations are overlapping because one variable—the variable coding the type of system—is part of both correlations being compared. As a side note, the fact that the correlations between the 10-item scales and the single-item scales reported in Table 7 are quite high is consistent with the assumption that both the two verbal scales and the two pictorial scales measure the same construct.

It is already obvious from the relative size of the correlations presented in Table 6 that the single-item Adjective Rating Scale reflects the difference in usability between the well-designed system and the poorly designed system at least as well as the 10-item System Usability Scale: At the descriptive level the single-item Adjective Rating Scale reflects the difference in usability even better than the 10-item System Usability Scale. A two-sided Steiger (1980) test for differences between dependent and overlapping correlations shows that the advantage of the Adjective Rating Scale over the System Usability Scale is statistically significant (Table 8). A similar result was obtained for the two pictorial scales. The relative size of the correlations presented in Table 6 shows that, at the descriptive level, the Pictorial Single-Item Usability Scale reflects the difference in usability between the systems even somewhat better than the 10-item Pictorial System Usability Scale. A two-sided Steiger (1980) test for differences between dependent and overlapping correlations shows that there is no statistically significant difference between the Pictorial Single-Item Usability Scale and the Pictorial System Usability Scale in how well these scales reflect the difference in usability between the well-designed system and the poorly designed system (Table 8).

Finally, to test whether there is a difference between the verbal and the pictorial perceived-usability scales in how well they reflect the difference in usability between the well-designed system and the poorly designed system, the original System Usability Scale was compared to the Pictorial System Usability Scale and the Adjective Rating Scale was compared to the Pictorial Single-Item Usability Scale using a two-sided Fisher (1925) test for independent correlations. These correlations are independent as disjunct groups of participants filled out the verbal and the pictorial scales. The results are reported in Table 9. They show that there was no statistically significant difference between the 10-item Pictorial System Usability Scale and the verbal 10-item System Usability Scale and also no statistically significant difference between the Pictorial Single-Item Usability Scale and the verbal single-item Adjective Rating Scale in how well these scales reflect the difference in usability between the two systems.

Table 7. Correlations between the 10-item perceived-usability scales and the single-item perceived-usability scales in Experiment 2.

	Adjective Rating Scale
System Usability Scale	.709 (.665 – .748)
	Pictorial Single-Item Usability Scale
Pictorial System Usability Scale	.727 (.685 – .764)

Note: The total score of the original 10-item System Usability Scale correlates positively with the score of the single-item Adjective Rating Scale (upper half of the table) and the total score of the Pictorial System Usability Scale correlates positively with the score of the Pictorial Single-Item Usability Scale (lower half of the table) across both systems. The 95 % confidence intervals are reported in parentheses.

Table 8. Results of the statistical test of whether a difference exists in how well the 10-item perceived-usability scales and the single-item perceived-usability scales reflect the usability difference between the well-designed system and the poorly designed system in Experiment 2.

	z	*p*
System Usability Scale vs. Adjective Rating Scale	2.450	.014
Pictorial System Usability Scale vs. Pictorial Single-Item Usability Scale	0.035	.972

Note: The verbal single-item scale is significantly better than its 10-item counterpart at reflecting the difference in usability between the well-designed system and the poorly designed system (upper row) and there is no statistically significant difference between the pictorial single-item scale and its 10-item counterpart in how well these scales reflect the difference in usability between the well-designed system and the poorly designed system (lower row), as shown by the Steiger (1980) test for differences between dependent and overlapping correlations implemented in *cocor* (Diedenhofen and Musch 2015).

Table 9. Results of the statistical test of whether a difference exists in how well the verbal perceived-usability scales as opposed to the pictorial perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system in Experiment 2.

	z	*p*
System Usability Scale vs. Pictorial System Usability Scale	0.493	.622
Adjective Rating Scale vs. Pictorial Single-Item Usability Scale	0.920	.357

Note: The 10-item (upper row) and single-item (lower row) pictorial perceived-usability scales reflect the difference in usability between the well-designed system and the poorly designed system as well as their verbal counterparts, as shown by the Fisher (1925) test for independent correlations.

Discussion

Consistent with the results of Experiment 1, the two single-item scales were descriptively better at reflecting the difference in usability between the well-designed system and the poorly designed system. In fact, the Adjective Rating Scale was even significantly better than the System Usability Scale at reflecting this difference.

In contrast to Experiment 1, the results of Experiment 2 do not lead to the conclusion that the verbal scales are better than their pictorial counterparts. Both the Pictorial System Usability Scale and the Pictorial Single-Item Usability Scale reflect the difference in usability between the well-designed system and the poorly designed system comparatively well (Table 6), and not significantly worse than their verbal counterparts (Table 9). From this pattern of results, we conclude that the advantage of the verbal over the pictorial scales observed in Experiment 1 cannot be robustly reproduced even with the large sample sizes that were used in the present series of experiments. Taken together, the findings of Experiments 1 and 2 suggest that using pictorial scales may be associated with somewhat reduced validity compared to the verbal scales but this reduction may be small and not reliable enough to argue against the use of pictorial perceived-usability scales, especially considering the potential of pictorial scales in helping respondents to overcome language barriers (Bradley and Lang 1994; Paunonen, Ashton, and Jackson 2001)

General discussion

Replicating and extending the findings by Gräve and Buchner (2024), the results of the statistical tests considered in conjunction with the descriptive data pattern obtained in Experiments 1 and 2 lead to the conclusion that the verbal single-item Adjective Rating Scale is at least as good as both the original and the positively worded 10-item System Usability Scale at reflecting differences in usability. This seems remarkable given that the System Usability Scale can currently be viewed as the gold standard for measuring perceived usability (cf. Bangor, Kortum, and Miller 2008).

Based on the present data, a parallel conclusion can now also be drawn for two pictorial perceived-usability scales. Specifically, the finding of no statistically significant difference between the Pictorial Single-Item Usability Scale and the Pictorial System Usability Scale in how well these scales reflect the difference between the well-designed system and the poorly designed system considered in conjunction with the fact that, at a descriptive level, the Pictorial Single-Item Usability Scale was even slightly better than the Pictorial System Usability Scale at reflecting this difference in both Experiment 1 and 2, we may conclude that the Pictorial Single-Item Usability Scale captures differences in perceived usability at least as well as the Pictorial System Usability Scale.

We may thus regard the Pictorial Single-Item Usability Scale as an easy-to-administer and efficient alternative to the 10-item Pictorial System Usability Scale, just as the verbal single-item Adjective Rating Scale is an easy-to-administer and efficient alternative to the verbal 10-item System Usability Scale. In both cases, the pragmatic advantages of single-item scales do not require any sacrifices in terms of validity.

In fact, there is a reason as to why single-item perceived-usability scales may be even somewhat better at representing the difference between systems differing in usability than their multi-item counterparts (cf. Gräve and Buchner 2024). The prompts provided by the verbal and the pictorial single item scales are relatively general and thus open to interpretation. As a consequence, users can flexibly focus on the most salient determinant of usability in a given context (e.g. the particularly bad legibility of the all-capitals negative polarity text in the poorly designed system used here, see Figure 1). This could maximise the difference between systems in terms of perceived-usability ratings. What is more, measuring perceived usability with a verbal or pictorial multi-item questionnaire may require participants to rate irrelevant or unknown aspects of an application. For instance, after having used an online store once, participants may not have a basis for answering verbal questions about how well integrated different functions of a system are (Item 5 of the System Usability Scale, see Table A1) or for giving an accurate rating on the corresponding pictorial item (5th row of the Pictorial System Usability Scale, see Table A4). In addition, having to respond to multiple verbal or pictorial items, many of which appear to measure essentially the same thing, may negatively affect the questionnaire experience (Baumgartner et al. 2021), thereby increasing the risk of undesirable response patterns such as giving random responses (Baumgartner et al. 2021, p. 2). If such effects exist, they may jeopardise the psychometric properties of a scale. However, whether this is really the case needs to be empirically investigated in future studies.

The conclusion that the verbal and pictorial single-item scales are as good as, or even somewhat better than, their multi-item counterparts at reflecting differences in usability fits with a broader literature suggesting that single-item scales provide valid

measurements in a large variety of domains (Table 1). We conclude from the present findings that perceived usability is another domain in which single-item scales yield valid measures. What do these domains have in common? From a theoretical point of view, it has been suggested that single-item scales yield valid measures of a construct whenever the to-be-measured construct and its attributes can be regarded to concern entities that are concrete and singular in the sense that (a) all raters understand which entity is being rated and (b) what is being rated is reasonably homogeneous (Rossiter 2002). In the case of perceived usability, it seems plausible that condition (a) is usually met. The same conclusion seems to hold for condition (b) given that the System Usability Scale can be treated as essentially unidimensional (Berkman and Karahoca 2016; Gräve and Buchner 2024; Kortum, Acemyan, and Oswald 2021; Lewis 2019; Lewis and Sauro 2017; Sauro 2018). In essence, then, the present results nicely fit the theoretical framework suggested by Rossiter (2002).

As clear as the present results are with respect to the validity of the single-item perceived-usability scales relative to their 10-item counterparts, it is less straightforward which conclusions can be drawn with respect to whether the verbal or the pictorial scales are to be preferred for perceived-usability measurement. In Experiment 1, the verbal scales turned out to be better than the pictorial scales at reflecting differences in usability but in Experiment 2 the pictorial scales were found to be as good as their verbal counterparts. The fact that there was no consistent advantage of the verbal over the pictorial scales even with the large sample sizes that were used in the present experiments suggests that the pictorial scales might be nearly as good as the verbal scales at reflecting differences in usability. Given this, it seems justified to use pictorial scales in situations in which non-verbal tools have specific advantages. For instance, pictorial scales should make it possible to include users with different levels of language proficiency. However, when the language proficiency of the respondents is not an issue, then these specific advantages of pictorial scales may not be so important. As a consequence, one might want to use verbal scales which, as suggested by the results of Experiment 1, could be somewhat better than their pictorial counterparts at reflecting differences in usability.

A side aspect of the present experiments is that the original System Usability Scale (Brooke 1996) was used in Experiment 2 instead of the positively worded version of the System Usability Scale (Sauro and Lewis 2011) that was used in Experiment 1. A quick look at the first column of Tables 2 and 6 shows that the two versions of the System Usability Scale yield essentially the same result. This strengthens the conclusions of Sauro and Lewis (2011) and Kortum, Acemyan, and Oswald (2021) that the two versions of the System Usability Scale lead to comparable results. Their conclusions thus hold true even for online environments in which the present experiments were implemented.

An obvious caveat when selecting the Adjective Rating Scale or the Pictorial Single-Item Usability Scale to measure perceived usability is that these scales provide only a single score representing the overall perception of a system's usability. Therefore, these single-item scales should only be seen as efficient alternatives to the total scores of their multi-item counterparts. Such overall measures of perceived usability are useful, for instance, when the goal is simply to track whether progressive changes to a system improve or worsen the usability of the system (see e.g. Bangor, Kortum, and Miller 2008, Figure 12). However, multi-item scales may offer additional information that could be helpful for identifying more specific problems (Baumgartner, Sonderegger, et al. 2019) or for locating usability weaknesses (Baumgartner, Frei, et al. 2019). For instance, if a system is perceived to be low in usability, the System Usability Scale may provide the information that the poor evaluation of the system's usability is mostly due to inconsistencies if the ratings in response to Item 6 of the System Usability Scale (Table A1) indicate a particularly negative view of the system's consistency. Such information would be particularly useful during formative evaluation. The Adjective Rating Scale cannot provide information at this level of detail and thus would be less useful in such circumstances.

However, for all cases in which information about the overall perceived usability of a system is sufficient, two key conclusions for practitioners can been drawn from the present results. First, the score derived from the verbal single-item Adjective Rating Scale is at least as good at reflecting perceived usability as the total score of the System Usability Scale. Second, the score derived from the Pictorial Single-Item Usability Scale is at least as good as the total score of the Pictorial System Usability Scale. Within the limits outlined in the previous paragraph, the verbal Adjective Rating Scale and the Pictorial Single-Item Usability Scale offer all the pragmatic benefits of single-item scales at no cost in terms of measuring overall differences in perceived usability.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors reported there is no funding associated with the work featured in this article.

ORCID

Elisa Gräve http://orcid.org/0000-0003-4653-7498
Raoul Bell http://orcid.org/0000-0002-0592-0362
Axel Buchner http://orcid.org/0000-0003-4529-3444

Data availability statement

The data of both experiments as well as screenshots of the online mobile phone contract purchasing systems are available at the project page of the Open Science Framework under https://osf.io/y5drh/

References

Ang, L., and M. Eisend. 2018. "Single versus Multiple Measurement of Attitudes: A Meta-Analysis of Advertising Studies Validates the Single-Item Measure Approach." *Journal of Advertising Research 58* (2): 218–227. doi:10.2501/JAR-2017-001.

Bangor, A., P. Kortum, and J. Miller. 2008. "An Empirical Evaluation of the System Usability Scale." *International Journal of Human-Computer Interaction 24* (6): 574–594. doi:10.1080/10447310802205776.

Bangor, A., P. Kortum, and J. Miller. 2009. "Determining What Individual SUS Scores Mean: adding an Adjective Rating Scale." *Journal of Usability Studies 4* (3): 114–123.

Baumgartner, J., N. Frei, M. Kleinke, J. Sauer, and A. Sonderegger. 2019. "Pictorial System Usability Scale (P-SUS): Developing an Instrument for Measuring Perceived Usability." CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK. doi:10.1145/3290605.3300299.

Baumgartner, J., N. Ruettgers, A. Hasler, A. Sonderegger, and J. Sauer. 2021. "Questionnaire Experience and the Hybrid System Usability Scale: Using a Novel Concept to Evaluate a New Instrument." *International Journal of Human-Computer Studies 147*: 102575. doi:10.1016/j.ijhcs.2020.102575.

Baumgartner, J., A. Sonderegger, and J. Sauer. 2019. "No Need to Read: Developing a Pictorial Single-Item Scale for Measuring Perceived Usability." *International Journal of Human-Computer Studies 122*: 78–89. doi:10.1016/j.ijhcs.2018.08.008.

Baxter, A. L., M. F. Watcha, W. V. Baxter, T. Leong, and M. M. Wyatt. 2011. "Development and Validation of a Pictorial Nausea Rating Scale for Children." *Pediatrics 127* (6): e1542–e1549. doi:10.1542/peds.2010-1410.

Bergkvist, L., and J. R. Rossiter. 2007. "The Predictive Validity of Multiple-Item versus Single-Item Measures of the Same Constructs." *Journal of Marketing Research 44* (2): 175–184. doi:10.1509/jmkr.44.2.175.

Bergkvist, L., and J. R. Rossiter. 2009. "Tailor-Made Single-Item Measures of Doubly Concrete Constructs." *International Journal of Advertising 28* (4): 607–621. doi:10.2501/S0265048709200783.

Berkman, M. I., and D. Karahoca. 2016. "Re-Assessing the Usability Metric for User Experience (UMUX) Scale." *Journal of Usability Studies 11* (3): 89–109.

Bieri, D., R. A. Reeve, G. D. Champion, L. Addicoat, and J. B. Ziegler. 1990. "The Faces Pain Scale for the Self-Assessment of the Severity of Pain Experienced by Children: Development, Initial Validation, and Preliminary Investigation for Ratio Scale Properties." *Pain 41* (2): 139–150. doi:10.1016/0304-3959(90)90018-9.

Bradley, M. M., and P. J. Lang. 1994. "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential." *Journal of Behavior Therapy and Experimental Psychiatry 25* (1): 49–59. doi:10.1016/0005-7916(94)90063-9.

Brooke, J. 1996. "SUS-A 'Quick and Dirty' Usability Scale." *Usability Evaluation in Industry 189* (194): 4–7. doi:10.1201/9781498710411-35.

Brooke, J. 2013. "SUS: A Retrospective." *Journal of Usability Studies 8* (2): 29–40.

Bruton, A. M., S. D. Mellalieu, and D. A. Shearer. 2016. "Validation of a Single-Item Stem for Collective Efficacy Measurement in Sports Teams." *International Journal of Sport and Exercise Psychology 14* (4): 383–401. doi:10.1080/1612197X.2015.1054853.

Buchner, A., and N. Baumgartner. 2007. "Text–Background Polarity Affects Performance Irrespective of Ambient Illumination and Colour Contrast." *Ergonomics 50* (7): 1036–1063. doi:10.1080/00140130701306413.

Cao, Y., C. Ritz, and R. Raad. 2013. How Much Longer To Go? The Influence of Waiting Time and Progress Indicators on Quality of Experience for Mobile Visual Search Applied to Print Media. *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Wörthersee, Austria. doi:10.1109/QoMEX.2013.6603220.

Cheung, F., and R. E. Lucas. 2014. "Assessing the Validity of Single-Item Life Satisfaction Measures: Results from Three Large Samples." *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation 23* (10): 2809–2818. doi:10.1007/s11136-014-0726-4.

Christophersen, T., and U. Konradt. 2011. "Reliability, Validity, and Sensitivity of a Single-Item Measure of Online Store Usability." *International Journal of Human-Computer Studies 69* (4): 269–280. doi:10.1016/j.ijhcs.2010.10.005.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Lawrence Earlbaum Associates. doi:10.4324/9780203771587.

de Boer, A. G. E. M., J. J. B. van Lanschot, P. F. M. Stalmeier, J. W. van Sandick, J. B. F. Hulscher, J. C. J. M. de Haes, and M. A. G. Sprangers. 2004. "Is a Single-Item Visual Analogue Scale as Valid, Reliable and Responsive as Multi-Item Scales in Measuring Quality of Life?" *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation 13* (2): 311–320. doi:10.1023/B:QURE.0000018499.64574.1f.

de Bruijn, L., D. E. J. Starreveld, M. Schaapveld, F. E. van Leeuwen, E. M. A. Bleiker, and N. E. Berentzen. 2022. "Single-Item Chronotype is Associated with Dim Light Melatonin Onset in Lymphoma Survivors with Fatigue." *Journal of Sleep Research 31* (5): e13577. doi:10.1111/jsr.13577.

DeSalvo, K. B., W. P. Fisher, K. Tran, N. Bloser, W. Merrill, and J. Peabody. 2006. "Assessing Measurement Properties of Two Single-Item General Health Measures." *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation 15* (2): 191–201. doi:10.1007/s11136-005-0887-2.

Diedenhofen, B., and J. Musch. 2015. "Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations." *PloS One 10* (3): e0121945. doi:10.1371/journal.pone.0121945.

Dolbier, C. L., J. A. Webster, K. T. McCalister, M. W. Mallon, and M. A. Steinhardt. 2005. "Reliability and Validity of a Single-Item Measure of Job Satisfaction." *American Journal of Health Promotion 19* (3): 194–198. doi:10.4278/0890-1171-19.3.194.

Eddy, C. L., K. C. Herman, and W. M. Reinke. 2019. "Single-Item Teacher Stress and Coping Measures: Concurrent and Predictive Validity and Sensitivity to Change." *Journal of School Psychology 76*: 17–32. doi:10.1016/j.jsp.2019.05.001.

Elo, A.-L., A. Leppänen, and A. Jahkola. 2003. "Validity of a Single-Item Measure of Stress Symptoms." *Scandinavian Journal of Work, Environment & Health 29* (6): 444–451. doi:10.5271/sjweh.752.

Fardell, J. E., V. Bray, M. L. Bell, B. Rabe, H. Dhillon, and J. L. Vardy. 2022. "Screening for Cognitive Symptoms among Cancer Patients during Chemotherapy: Sensitivity and Specificity of a Single Item Self-Report Cognitive Change Score." *Psycho-oncology 31* (8): 1294–1301. doi:10.1002/pon.5928.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang. 2009. "Statistical Power Analyses Using G* Power 3.1: Tests for Correlation and Regression Analyses." *Behavior Research Methods 41* (4): 1149–1160. doi:10.3758/BRM.41.4.1149.

Faul, F., E. Erdfelder, A.-G. Lang, and A. Buchner. 2007. "G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods 39* (2): 175–191. doi:10.3758/BF03193146.

Fedele, G., M. Fedriga, S. Zanuso, S. Mastrangelo, and F. Di Nocera. 2017. "Can User Experience Affect Buying Intention? A Case Study on the Evaluation of Exercise Equipment." Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2016 Annual Conference: Human Factors and Ergonomics Society Europe Chapter, Prague: Czech Republic. https://www.hfes-europe.org/wp-content/uploads/2016/11/Fedele2017.pdf.

Finstad, K. 2006. "The System Usability Scale and Non-Native English Speakers." *Journal of Usability Studies 1* (4): 185–188.

Finstad, K. 2010. "The Usability Metric for User Experience." *Interacting with Computers 22* (5): 323–327. doi:10.1016/j.intcom.2010.04.004.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. 1 ed. Edinburgh and London: Oliver and Boyd.

Gardner, D. G., L. L. Cummings, R. B. Dunham, and J. L. Pierce. 1998. "Single-Item versus Multiple-Item Measurement Scales: An Empirical Comparison." *Educational and Psychological Measurement 58* (6): 898–915. doi:10.1177/0013164498058006003.

Gogol, K., M. Brunner, T. Goetz, R. Martin, S. Ugen, U. Keller, A. Fischbach, and F. Preckel. 2014. "My Questionnaire is Too Long!" the Assessments of Motivational-Affective Constructs with Three-Item and Single-Item Measures." *Contemporary Educational Psychology 39* (3): 188–205. doi:10.1016/j.cedpsych.2014.04.002.

Graf, L., S. Mayer, and J. Landwehr. 2017. "Measuring Processing Fluency: 1 versus 5 Items." *Journal of Consumer Psychology 28* (3): 393–411. doi:10.1002/jcpy.1021.

Gräve, E., and A. Buchner. 2024. "Is Less Sometimes More? An Experimental Comparison of Four Measures of Perceived Usability." *Human Factors*. doi:10.1177/00187208241237862.

Gronier, G., and A. Baudet; Departement of Information Technology for Innovative Sciences (ITIS), Luxembourg Institute of Science and Technology (LIST), 5 av. des Hauts Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg. 2019. "Does Progress Bars' Behavior Influence the User Experience in Human-Computer Interaction." *Psychology and Cognitive Sciences – Open Journal 5* (1): 6–13. doi:10.17140/PCSOJ-5-144.

Hall, R. H., and P. Hanna. 2004. "The Impact of Web Page Text-Background Colour Combinations on Readability, Retention, Aesthetics and Behavioural Intention." *Behaviour & Information Technology 23* (3): 183–195. doi:10.1080/01449290410001669932.

Himmels, C., K. Omozik, O. Jarosch, and A. Buchner. 2021. Measuring User Experience in Automated Driving: Developing a Single-Item Measure . *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Leeds*, United Kingdom. doi:10.1145/3409118.3475135.

ISO-9241-210 2010. *Ergonomics of Human–System Interaction–Part 210: Human-Centred Design for Interactive Systems*. Geneva, Switzerland: International Standardization Organization (ISO).

ISO-9241/10. 1995. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) – Part 10: Dialogue Principles. Geneva, Switzerland: International Standardization Organization (ISO).

Jordan, J. S., and B. A. Turner. 2008. "The Feasibility of Single-Item Measures for Organizational Justice." *Measurement in Physical Education and Exercise Science 12* (4): 237–257. doi:10.1080/10913670802349790.

Jovanović, V., and M. Lazić. 2020. "Is Longer Always Better? A Comparison of the Validity of Single-Item versus Multiple-Item Measures of Life Satisfaction." *Applied Research in Quality of Life 15* (3): 675–692. doi:10.1007/s11482-018-9680-6.

Kortum, P., C. Z. Acemyan, and F. L. Oswald. 2021. "Is It Time to Go Positive? Assessing the Positively Worded System Usability Scale (SUS)." *Human Factors 63* (6): 987–998. doi:10.1177/0018720819881556.

Kunin, T. 1955. "The Construction of a New Type of Attitude Measure." *Personnel Psychology 8* (1): 65–77. doi:10.1111/j.1744-6570.1955.tb01189.x.

Kuo, W.-Y., C.-Y. Chen, M.-C. Chen, C.-M. Wang, Y.-L. Lin, and J. Wang. 2023. "Can Rehabilitation Adherence among Stroke Patients Be Measured Using a Single Item?" *Journal of Clinical Nursing 32* (5–6): 950–962. doi:10.1111/jocn.16544.

Kwon, H., and G. Trail. 2005. "The Feasibility of Single-Item Measures in Sport Loyalty Research." *Sport Management Review 8* (1): 69–89. doi:10.1016/s1441-3523(05)70033-4.

Leiner, D. J. 2022. SoSci Survey (Version 3.3.04) [Computer software]. https://www.soscisurvey.de.

Lewis, J. R. 2018. "Measuring Perceived Usability: The CSUQ, SUS, and UMUX." *International Journal of Human–Computer Interaction 34* (12): 1148–1156. doi:10.1080/10447318.2017.1418805.

Lewis, J. R. 2019. "Measuring Perceived Usability: SUS, UMUX, and CSUQ Ratings for Four Everyday Products." *International Journal of Human–Computer Interaction 35* (15): 1404–1419. doi:10.1080/10447318.2018.1533152.

Lewis, J. R., and J. Sauro. 2017. "Revisiting the Factor Structure of the System Usability Scale." *Journal of Usability Studies 12* (4): 183–192.

Lin, C.-C. 2003. "Effects of Contrast Ratio and Text Color on Visual Performance with TFT-LCD." *International Journal of Industrial Ergonomics 31* (2): 65–72. doi:10.1016/S0169-8141(02)00175-0.

Lohr, L. 2000. "Designing the Instructional Interface." *Computers in Human Behavior 16* (2): 161–182. doi:10.1016/S0747-5632(99)00057-6.

Lorish, C. D., and R. Maisiak. 1986. "The Face Scale: A Brief, Nonverbal Method for Assessing Patient Mood." *Arthritis & Rheumatism 29* (7): 906–909. doi:10.1002/art.1780290714.

Maldonado, C. C., A. J. Bentley, and D. Mitchell. 2004. "A Pictorial Sleepiness Scale Based on Cartoon Faces." *Sleep 27* (3): 541–548. doi:10.1093/sleep/27.3.541.

Mashek, D., L. W. Cannaday, and J. P. Tangney. 2007. "Inclusion of Community in Self Scale: A Single-Item Pictorial Measure of Community Connectedness." *Journal of Community Psychology 35* (2): 257–275. doi:10.1002/jcop.20146.

McKenzie, N., and I. Marks. 1999. "Quick Rating of Depressed Mood in Patients with Anxiety Disorders." *The British Journal of Psychiatry: The Journal of Mental Science 174* (3): 266–269. doi:10.1192/bjp.174.3.266.

Molich, R., and J. Nielsen. 1990. "Improving a Human-Computer Dialogue." *Communications of the ACM 33* (3): 338–348. doi:10.1145/77481.77486.

Nagy, M. S. 2002. "Using a Single-Item Approach to Measure Facet Job Satisfaction." *Journal of Occupational and Organizational Psychology 75* (1): 77–86. doi:10.1348/096317902167658.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science 349* (6251): aac4716. doi:10.1126/science.aac4716.

Oshagbemi, T. 1999. "Overall Job Satisfaction: How Good Are Single versus Multiple-Item Measures?" *Journal of Managerial Psychology 14* (5): 388–403. doi:10.1108/02683949910277148.

Pashutina, Y., S. Kastaun, E. Ratschen, L. Shahab, and D. Kotz. 2021. "External Validation of a Single-Item Scale to Measure Motivation to Stop Smoking: Findings from a Representative Population Survey (DEBRA Study)." *SUCHT 67* (4): 171–180. doi:10.1024/0939-5911/a000719.

Paunonen, S. V., M. C. Ashton, and D. N. Jackson. 2001. "Nonverbal Assessment of the Big Five Personality Factors." *European Journal of Personality 15* (1): 3–18. doi:10.1002/per.385.

Piepenbrock, C., S. Mayr, and A. Buchner. 2014. "Positive Display Polarity is Particularly Advantageous for Small Character Sizes: Implications for Display Design." *Human Factors 56* (5): 942–951. doi:10.1177/0018720813515509.

Piepenbrock, C., S. Mayr, I. Mund, and A. Buchner. 2013. "Positive Display Polarity is Advantageous for Both Younger and Older Adults." *Ergonomics 56* (7): 1116–1124. doi:10.1080/00140139.2013.790485.

Prümper, J. 1993. "Software-Evaluation Based upon ISO 9241 Part 10." In *Human Computer Interaction. VCHCI 1993*, edited by Grechenig & Tscheligi, 255–265. Vol. 733. Berlin, Heidelberg: Springer. doi:10.1007/3-540-57312-7_74.

Rammstedt, Beatrice, David J. Grüning, and Clemens M. Lechner. in press. "Measuring Growth Mindset: Validation of a Three-Item and a Single-Item Scale in Adolescents and Adults." *European Journal of Psychological Assessment* 40 (1): 84–95. doi:10.1027/1015-5759/a000735.

Robins, R. W., H. M. Hendin, and K. H. Trzesniewski. 2001. "Measuring Global Self-Esteem: Construct Validation of a Single-Item Measure and the Rosenberg Self-Esteem Scale." *Personality and Social Psychology Bulletin 27* (2): 151–161. doi:10.1177/0146167201272002.

Rossiter, J. R. 2002. "The C-OAR-SE Procedure for Scale Development in Marketing." *International Journal of Research in Marketing 19* (4): 305–335. doi:10.1016/S0167-8116(02)00097-6.

Sauro, J. 2018. Can you use a single item to predict SUS scores? *Measuring U.* https://measuringu.com/single-item-sus/

Sauro, J., and J. R. Lewis. 2011. When Designing Usability Questionnaires, Does it Hurt to be Positive? *Proceedings of the SIGCHI conference on human factors in computing systems,* Vancouver, BC, Canada. doi:10.1145/1978942.1979266.

Scapin, D. L., and J. C. Bastien. 1997. "Ergonomic Criteria for Evaluating the Ergonomic Quality of Interactive Systems." *Behaviour & Information Technology 16* (4-5): 220–231. doi:10.1080/014492997119806.

Seale, J. P., J. M. Boltri, S. Shellenberger, M. M. Velasquez, M. Cornelius, M. Guyinn, I. Okosun, and H. Sumner. 2006. "Primary Care Validation of a Single Screening Question for Drinkers." *Journal of Studies on Alcohol 67* (5): 778–784. doi:10.15288/jsa.2006.67.778.

Shamir, B., and R. Kark. 2004. "A Single-Item Graphic Scale for the Measurement of Organizational Identification." *Journal of Occupational and Organizational Psychology 77* (1): 115–123. doi:10.1348/096317904322915946.

Smith, A. B., M. Gao, M. Tran, M. Ftanou, S. Jegathees, V. Wu, M. Jefford, F. Lynch, H. M. Dhillon, J. Shaw, L. McDowell, A. White, C. Halloran, D. Wiesenfeld, and A. Bamgboje-Ayodele. 2023. "Evaluation of the Validity and Screening Performance of a Revised Single-Item Fear of Cancer Recurrence Screening Measure (Fcr-1r)." *Psycho-oncology 32* (6): 961–971. doi:10.1002/pon.6139.

Steiger, J. H. 1980. "Tests for Comparing Elements of a Correlation Matrix." *Psychological Bulletin 87* (2): 245–251. doi:10.1037/0033-2909.87.2.245.

Tinker, M. A. 1955. "Prolonged Reading Tasks in Visual Research." *Journal of Applied Psychology 39* (6): 444–446. doi:10.1037/h0041553.

Truby, H., and S. J. Paxton. 2002. "Development of the Children's Body Image Scale." *British Journal of Clinical Psychology 41* (2): 185–203. doi:10.1348/014466502163967.

Wanous, J. P., A. E. Reichers, and M. J. Hudy. 1997. "Overall Job Satisfaction: How Good Are Single-Item Measures?" *The Journal of Applied Psychology 82* (2): 247–252. doi:10.1037/0021-9010.82.2.247.

Weigelt, Oliver, Petra Gierer, Roman Prem, Michael Fellmann, Fabienne Lambusch, Katja Siestrup, Bernd Marcus, Thomas Franke, Sara Tsantidis, Miriam Golla, Claudia Wyss, and Johanna Blume. 2022. "Time to Recharge Batteries–Development and Validation of a Pictorial Scale of Human Energy." *European Journal of Work and Organizational Psychology 31* (5): 781–798. doi:10.1080/1359432X.2022.2050218.

Wong, S. M. Y., B. Y. H. Lam, C. S. M. Wong, H. P. Y. Lee, G. H. Y. Wong, S. S. Y. Lui, K. T. Chan, M. T. H. Wong, S. K. W. Chan, W. C. Chang, E. H. M. Lee, Y. N. Suen, C. L. M. Hui, and E. Y. H. Chen. 2021. "Measuring Subjective Stress among Young People in Hong Kong: Validation and Predictive Utility of the Single-Item Subjective Level of Stress (SLS-1) in Epidemiological and Longitudinal Community Samples." *Epidemiology and Psychiatric Sciences 30*: e61. doi:10.1017/S2045796021000445.

Zhang, J. W., R. T. Howell, S. Chen, A. R. Goold, B. Bilgin, W. J. Chai, and T. Ramis. 2022. "I Have High Self-Compassion': A Face-Valid Single-Item Self-Compassion Scale for Resource-Limited Research Contexts." *Clinical Psychology & Psychotherapy* 29 (4): 1463–1474. doi:10.1002/cpp.2714.

Appendix A

Table A1. Items of the original System Usability Scale.

	Items
1.	I think that I would like to use this system frequently.
2.	I found the system unnecessarily complex.
3.	I thought the system was easy to use.
4.	I think that I would need the support of a technical person to be able to use this system.
5.	I found the various functions in the system were well integrated.
6.	I thought there was too much inconsistency in this system.
7.	I would imagine that most people would learn to use this system very quickly.
8.	I found the system very awkward to use.
9.	I felt very confident using the system.
10.	I needed to learn a lot of things before I could get going with this system.

Note: The version of the original System Usability Scale is shown in which "cumbersome" has been replaced by "awkward" in item 8. Statements are evaluated on a 5-point Likert scale of strength of agreement. For the System Usability Scale total score, responses on a scale from 1 to 5 are coded as values between 0 and 4 for positively worded items and from 4 to 0 for negatively worded items, summed up and the sum is multiplied by 2.5. Higher values indicate better perceived usability.

Table A2. Items of the positively worded System Usability Scale.

	Items
1.	I think that I would like to use this system frequently.
2.	I found the system to be simple.
3.	I thought the system was easy to use.
4.	I think I could use the system without the support of a technical person.
5.	I found the various functions in the system were well integrated.
6.	I thought there was a lot of consistency in the system.
7.	I would imagine that most people would learn to use this system very quickly.
8.	I found the system very intuitive.
9.	I felt very confident using the system.
10.	I could use the system without having to learn anything new.

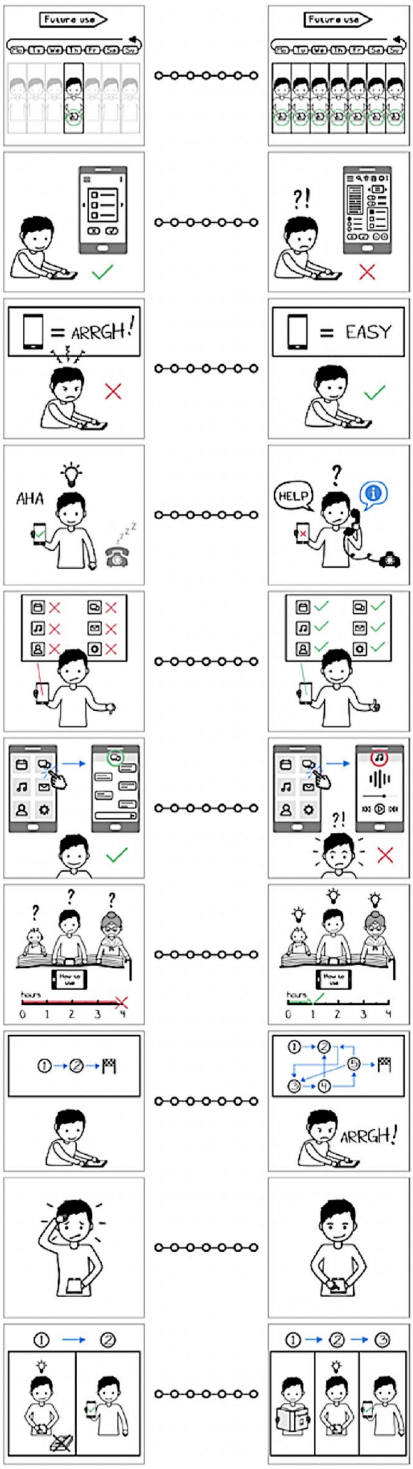
Note: The positive System Usability Scale statements are evaluated on a 5-point Likert scale of strength of agreement. For the System Usability Scale total score, responses on a scale from 1 to 5 are coded as values between 0 and 4, summed up and the sum is multiplied by 2.5. Higher values indicate better perceived usability.

Table A3. The Adjective Rating Scale.

Overall, I would rate the user-friendliness of this product as:
Worst imaginable
Awful
Poor
Ok
Good
Excellent
Best Imaginable

Note: Perceived usability is evaluated using a 7-point, adjective-anchored Likert scale. Responses on a scale from 1 (worst imaginable) to 7 (best imaginable) are coded as values between 0 and 6, divided by 6 and multiplied by 100. Higher values indicate better perceived usability.

Table A4. Items of the Pictorial System Usability Scale.



From: "Pictorial System Usability Scale (P-SUS): Developing an instrument for measuring perceived usability" by J. Baumgartner, N. Frei, M. Kleinke, J. Sauer & A. Sonderegger, 2019, CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK (https://doi.org/10.1145/3290605.3300299). Copyright 2019 by the author(s). Reprinted with permission.

Note: For the Pictorial System Usability Scale total score, responses from left to right are coded as values between 0 and 6 for odd items and as values between 6 and 0 for even items, summed up and the sum is multiplied by 5/3. Higher values indicate better perceived usability.

Table A5. The Pictorial Single-Item Usability Scale.

From: "No need to read: Developing a pictorial single-item scale for measuring perceived usability" by J. Baumgartner, A. Sonderegger & J. Sauer, 2019, International Journal of Human-Computer Studies, 122, p. 8 (https://doi.org/10.1016/j.ijhcs.2018.08.008). Copyright 2019 by Elsevier. Reprinted with permission.

Note: Responses to the Pictorial Single-Item Usability Scale are collected using a bipolar 9-point, icon-anchored Likert scale. Responses on a scale from –4 to 4 are coded as values between 0 and 8, divided by 8 and multiplied by 100. Higher values indicate better perceived usability.

Erklärung über den Eigenanteil an den in der Dissertation enthaltenen Einzelarbeiten

Meine Dissertationsschrift umfasst zwei Fachartikel mit insgesamt vier Experimenten, die in zwei verschiedenen wissenschaftlichen Fachzeitschriften publiziert wurden. Im Folgenden ist für jeden der Fachartikel mein eigener Anteil sowie der Anteil der einzelnen Koautoren aufgeführt.

Eigenanteil an Einzelarbeit 1

Publikation:

Gräve, E., & Buchner, A. (2024). Is less sometimes more? An experimental comparison of four measures of perceived usability. *Human Factors*. https://doi.org/10.1177/00187208241237862

Planung: Das experimentelle Design habe ich mit der Unterstützung von Axel Buchner erstellt.

Umsetzung: Die Experimente und Auswertungsprogramme programmierte ich selbstständig. Axel Buchner überprüfte die Korrektheit und äußerte konstruktive Kritik und Verbesserungsvorschläge. Die Datenerhebung führte ich eigenständig durch.

Auswertung: Die statistischen Analysen führte ich mittels SPSS eigenständig durch. Axel Buchner überprüfte diese auf Korrektheit. Bei der Durchführung einer Parallelanalyse (Horn, 1965) im Programm R wurde ich von Nils Brandenburg und Axel Buchner unterstützt, die mir den Ablauf erklärten und die Ergebnisse vermittelten.

Manuskript: Das Manuskript habe ich eigenständig verfasst, von der umfassenden Literaturrecherche bis hin zur finalen Ausarbeitung. Die Abbildungen und Tabellen erstellte ich eigenständig. Axel Buchner gab mir konstruktive Verbesserungsvorschläge, die ich nach sorgfältiger Prüfung in das Manuskript einarbeitete. Den Revisionsprozess bei der wissenschaftlichen Fachzeitschrift koordinierte ich selbstständig. Während dieses Prozesses nahm ich, unterstützt von

Axel Buchner, die notwendigen Revisionen vor. Die finale Version des Manuskripts erstellte ich eigenständig.

Eigenanteil an Einzelarbeit 2

Publikation:

Gräve, E., Bell, R., & Buchner, A. (2024). Verbal and pictorial single-item scales are as good as their 10-item counterparts for measuring perceived usability. *Ergonomics.* https://doi.org/10.1080/00140139.2024.2371061

Planung: Das experimentelle Design habe ich mit der Unterstützung von Axel Buchner und Raoul Bell erstellt.

Umsetzung: Die Experimente und Auswertungsprogramme programmierte ich selbstständig. Axel Buchner und Raoul Bell überprüften die Korrektheit und äußerten konstruktive Kritik und Verbesserungsvorschläge. Die Datenerhebung führte ich eigenständig durch.

Auswertung: Die statistischen Analysen führte ich mittels SPSS eigenständig durch. Axel Buchner und Raoul Bell überprüften diese auf Korrektheit.

Manuskript: Das Manuskript habe ich eigenständig verfasst, von der umfassenden Literaturrecherche bis hin zur finalen Ausarbeitung. Die Abbildungen und Tabellen erstellte ich eigenständig. Axel Buchner und Raoul Bell gaben mir konstruktive Verbesserungsvorschläge, die ich nach sorgfältiger Prüfung in das Manuskript einarbeitete. Den Revisionsprozess bei der wissenschaftlichen Fachzeitschrift koordinierte ich selbstständig. Während dieses Prozesses nahm ich, unterstützt von Axel Buchner und Raoul Bell, die notwendigen Revisionen vor. Die finale Version des Manuskripts erstellte ich eigenständig.

Erklärung an Eides statt

Hiermit versichere ich an Eides statt, dass ich die Dissertation mit dem Titel „Messung wahrgenommener Benutzbarkeit: Vergleich von verbalen und piktoralen Ein-Item- und Multi-Item-Fragebogen“ selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt habe.

Ich versichere insbesondere:

(1) Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.

(2) Alle wörtlich oder dem Sinn nach aus anderen Texten entnommenen Stellen habe ich als solche kenntlich gemacht; dies gilt für gedruckte Texte ebenso wie für elektronische Ressourcen.

(3) Die Arbeit habe ich in der vorliegenden oder einer modifizierten Form noch nicht als Dissertation vorgelegt – sei es an der Heinrich-Heine-Universität oder an einer anderen Universität.

Datum: 29. Oktober 2024

Name: Elisa Katharina Gräve

Unterschrift: