Figl, K.(2010). Deutschsprachige Fragebögen zur Usability-Evaluation im Vergleich. Zeitschrift für Arbeitswissenschaft. 2010/4.

Deutschsprachige Fragebögen zur Usability-Evaluation im Vergleich

Autorenbeschreibung: Dr. MMag. Kathrin Figl, Institut für Wirtschaftsinformatik und Neue Medien,WU-Wirtschaftsuniversität Wien, UZAII, Augasse 2-6, A-1090 Wien, Österreich

Titel [Englisch]: German Questionnaires for Evaluating Usability

Schlüsselwörter [Deutsch]: Usability, EN ISO 9241-110, Software-Ergonomie, Software-Evaluation, Fragebögen, Isometrics, Isonorm 9241/10

Schlüsselwörter [Englisch]: Usability, EN ISO 9241-110, Human Computer Interaction, Software Evaluation, Questionnaires, Isometrics, Isonorm 9241/10

Zusammenfassung [Deutsch]: Für die Konstruktion gebrauchstauglicher Anwendungssysteme ist eine exakte Evaluierung der Usability eine wertvolle Unterstützung. Zu diesem Zweck werden in der Praxis häufig Usability-Fragebögen herangezogen. Im deutschen Sprachraum sind die beiden Fragebögen Isonorm 9241/10 und Isometrics, die beide Software gemäß der EN ISO 9241–110 evaluieren, weit verbreitet. Die vorliegende Studie widmete sich einem Vergleich dieser beiden Fragebögen hinsichtlich testtheoretischer Gütekriterien. Im Rahmen eines experimentellen Designs wurden die beiden Fragebögen eingesetzt um die Usability von zwei Standard-Softwarepaketen zu bewerten. Hinsichtlich der inhaltlichen Validität der Fragebögen zeigten die Ergebnisse eine hohe Übereinstimmung der Usability-Messung der beiden Fragebögen. Auch weitere testtheoretische Analysen lieferten eine ähnliche Qualitätsbeurteilung beider Fragebögen, weshalb sie aus diesem Blickwinkel gleichermaßen für Forschung und Praxis empfohlen werden können.

Summary [English]: Software usability is a highly relevant issue in the ergonomics of information systems, as these systems are prevalent in today's workplace. Accurate evaluation methods are crucial for the construction of usable software that helps users to achieve their goals in an effective, efficient and satisfying way. For this purpose, usability engineers frequently employ usability questionnaires in practice. Usability questionnaires are economic in their administration and scoring, and they allow for accurate subjective assessments of users with a standardized method. In German, there exist two widely used questionnaires, Isonorm 9241/10 and Isometrics, which both evaluate software in accordance with the 7 criteria of the EN ISO 9241-110 (suitability for the task, suitability for learning, suitability for individualization, conformity with user expectations, self-descriptiveness, controllability and error tolerance). Both questionnaires offer scales for each usability criterion, but they differ in the amount of items (35 items in the Isonorm questionnaire, in comparison to 75 items in the Isometrics questionnaire), the wording of the items and the answering format. Although previous research has analyzed these questionnaires separately, little research has been devoted to improving our understanding of how these questionnaires differ and whether their results are comparable. To fill this gap, our study seeks to compare these two questionnaires regarding test theoretical criteria. Additionally, we investigate whether users prefer one questionnaire to another. Concerning test theoretical criteria, differences in objectivity are unlikely, as the administration and scoring of both questionnaires are detailed in the manuals.

Still, differences in validity and reliability need further inquiry. To this end, we conducted an experiment in which we used the two questionnaires to evaluate two standard software packages. In the controlled two-group design, a total of 50 participants took part. The experiment was conducted via laptops, and the total duration was an average of 50 minutes per user. First, we acquainted the participants with one of the two statistical software packages in a usability test comprised of 14 tasks. To complete those tasks, the participants had to carry out different statistical procedures on a small data set, to eliminate input errors and to use the help functions. The tasks and instructions were constructed in a way to give participants an overview of as many different facets of the software as possible. Then, participants filled out both usability questionnaires. We used different scramblings to avoid order effects. Regarding the content validity of the questionnaires, the results showed a high agreement between usability measurements of the two questionnaires. According to an analysis of variance, the factor "questionnaire" did not have a significant influence on the usability measurement. For instance, results from both questionnaires showed that users rated one of the software packages better on the scale for controllability. Additionally, the scales of the questionnaires showed medium to high correlations. In general, participants who spent more time on the tasks rated the usability lower, which demonstrated the criterion-related validity of both questionnaires. Moreover, both questionnaires offer satisfying reliability, and there were only a few negatively worded items that would heighten reliability if excluded. According to individual preferences, half of the users thought that Isonorm was more enjoyable to answer, the other half preferred Isometrics; therefore, we cannot decisively conclude which questionnaire was more popular among users. However, participants needed more extensive knowledge of the software system to answer some items of Isometrics than to answer the items of Isonorm. As a typical example, to answer some items in the Isometrics questionnaire of the scale of "conformity with user expectations", the users needed long-term experience with the software. In summary, our test-theoretical analyses indicate that the use of both questionnaires leads to similar results. Because we used only two different software packages in this experiment, one must generalize our findings with caution, and further research is necessary to validate our results. A possible direction for future inquiry emerges from our study. Since our data showed high intercorrelations between different usability scales in both questionnaires, a study that uses higher sample sizes might examine whether different usability criteria are measurable independently from each other or whether they can be explained with one general usability factor. In conclusion, based on our findings, both questionnaires can be recommended equally for research and practice from a test theoretical point of view.

Praktische Relevanz [Deutsch]: Da in der Praxis und der Forschung unterschiedliche Usability-Fragebögen eingesetzt werden, stellt sich die Frage, inwiefern und ob Evaluationsergebnisse auf Basis verschiedener Usability-Fragebögen vergleichbar sind. Die vorliegende Arbeit versucht dieser Problemstellung auf den Grund zu gehen, indem sie die beiden am meisten verbreiteten Usability-Fragebögen Isonorm 9241/10 und Isometrics vergleicht. Für die Praxis und für Anwender mit geringem testtheoretischem Hintergrund bietet ein Vergleich der Qualität von Usability-Fragebögen eine wichtige Entscheidungs- und Orientierungshilfe für deren Auswahl.

Practical Relevance [English]: The fact that different usability questionnaires are used in practice as well as in research begs the question of whether and to what extent evaluation results stemming from different questionnaires are comparable. The present work seeks to investigate this issue by comparing two widely used German usability questionnaires, Isonorm 9241/10 and Isometrics. From a practical perspective, the knowledge gained as a

result of the present study can provide advice for practitioners and users, who have low test-theoretical backgrounds, to choose a questionnaire.

Anmerkung: Der vorliegende Artikel ist eine gänzlich überarbeitete und erweiterte Version des Konferenzbeitrages „Figl, K. (2009). Usability-Fragebögen im Vergleich. Tagungsband Mensch & Computer. Oldenbourg: Berlin, 143-152."

1 Einleitung

Aufgrund der immer stärkeren Durchdringung der Informationstechnologie der Arbeitswelt, ist Usability von Software und deren Einfluss auf das Arbeitsleben für die Arbeitswissenschaft von zentralem Interesse. Usability ist ein wesentlicher Aspekt der Qualität von Anwendungssystemen und ist für die Arbeitsproduktivität beim Softwareeinsatz, die Wirtschaftlichkeit einer Softwareinvestition und die Wettbewerbsfähigkeit von Produkten relevant. Usability bzw. „ease of use" beeinflusst die wahrgenommene Nützlichkeit und die Absicht, ein System zu verwenden (Davis 1989). Neben der Informationsqualität ist sie ein wichtiger Bestandteil der Systemqualität, welche ein wesentlicher Faktor für die Systemverwendung und Nutzerzufriedenheit ist (Delone and McLean 2003). Daher kommt der Usability auch im Rahmen der Erfolgsmessung von Informationssystemen (siehe Urbach, Smolnik et al. 2009 für eine Übersicht) eine wichtige Bedeutung zu.

Um die Usability von Software zu messen, sind Usability-Fragebögen eine in Forschung und Praxis sehr häufig eingesetzte Evaluationsmethode. Usability-Fragebögen stellen im Vergleich zu anderen Methoden eine kostengünstige Möglichkeit, Usability zu erheben dar, die gleichzeitig eine hohe Durchführungs- und Auswertungsökonomie bietet. Mit ihrer Hilfe können subjektive Bewertungen der Benutzer in standardisierter Form erfasst sowie Usability-Probleme aufgedeckt werden. Liegen für einen Fragebogen auch Normierungswerte vor, so bieten sie außerdem die Möglichkeit, Softwaresysteme auf einfachem Wege mit einander zu vergleichen. Eine aktuelle Metaanalyse der Verwendung von Usability-Evaluationsmethoden in der Forschung zeigte typische methodische Schwächen beim Einsatz von Usability-Fragebögen auf (Hornbæk 2006). Einerseits wird in Usability-Studien zu selten über Validität und Reliabilität eingesetzter Skalen berichtet, andererseits werden oftmals eigens neu konstruierte Fragebogen-Items eingesetzt, obwohl standardisierte Usability-Fragebögen und Items bereits verfügbar wären. Die Vorgabe standardisierter Usability-Fragebögen würde z.B. den Vergleich zwischen Ergebnissen verschiedener Usability-Studien erleichtern. Diese Ergebnisse weisen auf die Wichtigkeit und Relevanz hin, sowohl Skalen standardisierter Fragebögen in der Forschung zu verwenden, als auch Qualitätskriterien bei häufig eingesetzten Fragebögen kritisch zu hinterfragen.

Im deutschen Sprachraum wurde in den letzten Jahren eine Reihe von Usability-Fragebögen veröffentlicht (siehe Prümper and Anft 1993; Hamborg, Willumeit et al. 1996; Dzida, Hofmann et al. 2000). Da in der Praxis und der Forschung unterschiedliche Usability-Fragebögen eingesetzt werden, stellt sich die Frage, inwiefern und ob Evaluationsergebnisse auf Basis verschiedener Usability-Fragebögen vergleichbar sind. Die vorliegende Arbeit versucht dieser Problemstellung auf den Grund zu gehen, indem sie die beiden am meisten verbreiteten Usability-Fragebögen Isonorm 9241/10 (im Folgenden mit Isonorm abgekürzt) und Isometrics vergleicht.

Ein Überblick über bereits vorhandene Forschungsergebnisse zu deutschen Usability-Fragebögen zeigt, dass, zwar einzelne Studien existieren, die diese hinsichtlich testtheoretischer Gütekriterien untersuchen (siehe z.B. Willumeit, Gediga et al. 1995;

Prümper 1997; Gediga, Hamborg et al. 1999), jedoch noch keine, in denen Usability-Fragebögen, deren Qualität und gelieferte Ergebnisse direkt verglichen werden. Aus diesem Grund ergibt sich hier ein wissenschaftliches Desiderat. Auch für die Praxis und für Anwender mit geringem testtheoretischem Hintergrund bietet ein Vergleich der Qualität von Usability-Fragebögen eine wichtige Entscheidungs- und Orientierungshilfe für deren Auswahl angesichts der stetig wachsenden Menge an Gestaltungsregeln und Usability-Fragebögen.

Der Artikel ist wie folgt aufgebaut: Nach einer allgemeinen Einführung in Usability mit Schwerpunkt auf der EN ISO 9241-110 wird ein Überblick über Usability-Fragebögen und Kriterien für deren Qualitätsbeurteilung gegeben. Dann werden das Design der empirischen Untersuchung, welche die Usability-Fragebögen Isonorm und Isometrics im Rahmen eines experimentellen Designs vergleicht, und deren Ergebnisse dargestellt. Der Artikel schließt mit einer Diskussion der wissenschaftlichen Ergebnisse sowie Handlungsempfehlungen für die Praxis ab.

2 Usability

Usability, auf Deutsch Gebrauchstauglichkeit, beschäftigt sich als Teilgebiet des Software-Engineerings mit Theorien und Methoden zur Konzeption, Realisierung und Testung von benutzergerechten Anwendungssystemen (Herczeg 2005). In der EN ISO 9241-11 wird Usability von Software definiert als das *„Ausmaß, in dem es von einem bestimmten Benutzer verwendet werden kann, um bestimmte Ziele in einem bestimmten Kontext effektiv, effizient und zufriedenstellend zu erreichen.*“ (Europäisches Komitee für Normung 1995). Die Begriffe „Effektivität“, „Effizienz“ und „Zufriedenstellung“ werden folgendermaßen definiert:

- **Effektivität:** Genauigkeit und Vollständigkeit, mit der Benutzer ein bestimmtes Ziel erreichen.
- **Effizienz:** Der im Verhältnis zur Genauigkeit und Vollständigkeit eingesetzte Aufwand, mit dem Benutzer ein bestimmtes Ziel erreichen.
- **Zufriedenstellung:** Freiheit von Beeinträchtigungen und positive Einstellungen gegenüber der Nutzung des Produkts.

Die Wichtigkeit verschiedener Usability-Eigenschaften variiert je nach Produkt. Für Produkte, die der Benutzer tagtäglich anwendet, wie z.B. Software für Callcenter, ist Effizienz unabdingbar (Nielsen 1996); für die meisten Webseiten ist z.B. eine schnelle Erlernbarkeit wichtig, da Benutzer diese oftmals nur wenige Male verwenden.

2.1 EN ISO 9241-110

Die EN ISO 9241-110 (Europäisches Komitee für Normung 2006) behandelt die ergonomische Gestaltung von interaktiven Systemen und sie formuliert 7 Prinzipien für die Beschreibung, das Design und die Evaluation der Dialoggestaltung von Software: Aufgabenangemessenheit, Selbstbeschreibungsfähigkeit, Steuerbarkeit, Erwartungskonformität, Fehlerrobustheit, Individualisierbarkeit, Lernförderlichkeit. Sie ersetzte 1996 die EN ISO Norm 9241-10 (Europäisches Komitee für Normung 1995), die sich inhaltlich stark auf Software im Bürokontext fokussiert hatte. Generell ist sie ein Teil des europäischen Standards EN ISO 9241 „Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten“ welcher abgeleitet von den internationalen ISO Standards vom europäischen Komitee für Normung herausgegeben wurde. Die EN ISO 9241 besteht aus 17 Teilen; die Teile 14-17 bieten spezielle Anleitungen für die Anwendung von Dialogen (z.B. Dialogführung mit Menüs, Kommandosprachen, Manipulation oder Bildschirmformularen),

und sie beschreiben unter anderem auch Anforderungen an Arbeitsaufgaben, Tastaturen und andere Eingabegeräte, zudem Anforderungen bezüglich der Arbeitsplatzgestaltung, Arbeitsumgebung und Farbdarstellung.

Die Prinzipien der EN ISO 9241-110 sind als voneinander abhängig zu verstehen, und manchmal wird es notwendig sein, Vorteile eines Grundsatzes gegenüber einem anderen abzuwägen (Europäisches Komitee für Normung 1995). Die konkrete Umsetzung der Grundsätze sollte sich an Merkmalen zukünftiger Benutzer, Arbeitsaufgaben, Arbeitsumgebung und eingesetzter Dialogtechnik orientieren. Wie auch in anderen für die Mensch-Computer-Interaktion relevanten Normen werden darin generelle Prinzipien definiert, die eine gute Hilfestellung für die Praxis bieten, ohne genaue Details für die Umsetzung von Interfaces vorzuschreiben (Bevan 2001).

Im Folgenden wird auf die 7 Grundsätze der EN ISO 9241-110 (Europäisches Komitee für Normung 2006) im einzelnen eingegangen.

1. **Aufgabenangemessenheit**: Ziele sollen auf einfachem und direktem Weg erreicht werden können, ohne dass der Benutzer zusätzlich, z.B. durch komplizierte Bedienung, belastet wird, so daß er seine Aufgabe effektiv und effizient erledigen kann.
2. **Selbstbeschreibungsfähigkeit**: Es soll ohne zusätzliche Beschriftungen, Erklärungen, Legenden für den Benutzer erkennbar sein, worum es sich bei einer Anzeige, einem Interaktionselement oder einer Eingabeaufforderung handelt.
3. **Steuerbarkeit**: Steuerbarkeit eines Dialoges ist gegeben, wenn der Benutzer den Dialogablauf starten und seine Richtung und Geschwindigkeit beeinflussen kann, um sein Ziel zu erreichen. Es soll z.B. möglich sein, einen Dialog zu unterbrechen oder Dialogschritte rückgängig zu machen.
4. **Erwartungskonformität**: Ein System sollte sich so verhalten, wie es ein Benutzer auf Grund von Vorerfahrungen erwartet und wie es dem Arbeitsgebiet und allgemeinen Konventionen entspricht (z.B. F1 - Taste für Hilfefunktion, Doppelklick dafür, um ein Anwendungsprogramm zu starten). Dialogverhalten und Informationsdarstellung sollten innerhalb eines Dialogsystems einheitlich sein.
5. **Fehlertoleranz**: Fehlertoleranz bedeutet, dass der Benutzer trotz unvollständiger oder fehlerhafter Eingabe mit minimalem Korrekturaufwand sein Ziel erreichen kann.
6. **Individualisierbarkeit**: Beispiele für Individualisierbarkeit wären die Möglichkeit der Verwendung größerer Schriftzeichen für Benutzer mit Sehbehinderung oder unterschiedlicher Tastenbelegung in verschiedenen Sprachräumen.
7. **Lernförderlichkeit**: Lernförderlichkeit kann durch Hilfefunktionen in jeglicher Form, wie z.B. Wizards, kontextsensitive Hilfe oder Handbücher unterstützt werden.

2.2 Evaluation der Software - Usability

In der Software-Ergonomie können drei grundsätzliche Evaluationsfragen unterscheiden werden (Gediga and Hamborg 2002):

- *„**Which is better?**“*: Bei dieser Evaluationsfragestellung werden alternative Softwaresysteme oder Designalternativen miteinander verglichen, um das am besten geeignete auszuwählen.

- *„**How good?**“*: Dabei wird untersucht, in welcher Ausprägung vorab definierte Usability-Ziele oder -Normen erreicht werden, um ein Softwareprogramm zu bewerten.
- *„**Why bad?**“*: Dieses Evaluationsziel beschäftigt sich damit, Schwachstellen und Gestaltungsvorschläge für eine Weiterentwicklung des Systems zu finden.

Prinzipiell gibt es sowohl aus der Hersteller- als auch der Käufersicht unterschiedliche Anlässe zur Evaluierung von Software (Zülch and Stowasser 2002). Für einen Softwarehersteller ist es z.B. sinnvoll, die Usability des Produktes im Rahmen der Softwareentwicklung zu evaluieren, um diese laufend zu verbessern und die Qualität zu steigern. Dabei ist es wichtig, Usability-Tests möglichst früh im Prozess einzusetzen, um Fehler früh erkennen zu können und potentielle Anwender frühzeitig einzubeziehen. Aus der Sicht der Käufer mag es z.B. bei größeren Softwareanschaffungen sinnvoll sein, im Auswahlprozess konkurrierende Produkte nicht nur gemäß funktionalen Anforderungen, sondern auch hinsichtlich der Usability zu evaluieren.

Hinsichtlich Evaluationsmethoden für Software-Usability liegen unterschiedliche Taxonomien vor. Gediga and Hamborg (2002) unterscheiden prädiktive von deskriptiven Evaluationsmethoden. Prädiktive Evaluationsmethoden zielen auf Identifikation von Schwächen und geben Gestaltungsempfehlungen ab. Zu ihnen zählen z.B. der „Walkthrough“, die Experteninspektion oder Gruppendiskussionen. Sie sind weniger aufwändig, benötigen weniger Information und sind früher einsetzbar als die deskriptiven Evaluationsmethoden. Letztere hingegen dienen dazu, den Status einer Software zu beschreiben und Benutzbarkeitsprobleme zu entdecken. Dazu zählen die verhaltensbasierten Evaluationsmethoden wie z.B. Verhaltensbeobachtung oder „lautes Denken“, die meinungsbasierten Methoden wie z.B. Fragebögen und auch Benutzbarkeitstests, in denen meist eine Mischung aus verhaltensbasierten und meinungsbasierten Methoden verwendet wird. Desweiteren können z.B. formale Evaluationsmethoden, bei denen Gestaltungslösungen von Experten beurteilt oder Simulationsmodelle eingesetzt werden, von empirischen unterschieden werden, bei denen Benutzer involviert sind (Hüttner, Wandke et al. 1995).

Im Rahmen empirischer Evaluationsmethoden wird oftmals die Interaktion eines Benutzers mit der Software beobachtet, um Probleme zu entdecken. Dazu werden Methoden wie Beobachtung (Hüttner, Wandke et al. 1995), Log-File Protokolle und Lautes Denken („Thinking aloud“) (Rauterberg 1992) verwendet. Sie können erst relativ spät in der Softwareentwicklung eingesetzt werden, da es möglich sein muss, dass Benutzer mit bereits lauffähigen Prototypen interagieren (Gediga and Hamborg 2002). Durch die Beobachtung können nicht nur qualitative, sondern auch quantitative Daten wie Bearbeitungszeiten, Fehlerraten, Art und Anzahl der genutzten Hilfen, Reaktionszeiten und Abfolgen von Benutzeraktionen erhoben werden (Hüttner, Wandke et al. 1995). Die Beobachtung wird oftmals auch mit der Analyse von Log-File Protokollen kombiniert, die Benutzereingaben und Systemmeldungen registrieren. Im Gegensatz zur Sammlung von Daten über das „äußere Verhalten“ mittels Videoaufzeichnungen, Beobachtungs- und Log-File Protokollen können im Rahmen von Methoden - wie Usability-Fragebögen, Interviews und Lautes Denken - auch Daten über innere Kognitionen sowie Emotionen und Einstellungen erfasst werden (Hüttner, Wandke et al. 1995). Beim Lauten Denken werden Benutzer gebeten, alle Gedanken und Überlegungen, die im Zusammenhang mit der Aufgabenlösung stehen, zu verbalisieren. Somit können handlungsbegleitende Kognitionen und Emotionen und Problemlösungsprozesse in qualitativer Weise erfasst werden.

In einer Meta-Studie von Hornbaek (2006) über aktuell verwendete Methoden in der Usability-Forschung zeigte sich, dass Genauigkeit der Aufgabenerreichung und Fehlerrate die am häufigsten verwendeten Effektivitätsmaßzahlen waren; die Zeitdauer für die Aufgabenbearbeitung und die Analyse von Lösungsmustern wurden besonders oft zur Effizienzmessung eingesetzt. Neben der objektiven Messung von Effektivität und Effizienz ist es auch wichtig, die subjektive Zufriedenheit der Nutzer zu messen (Hornbæk 2006), worauf der nächste Abschnitt näher eingeht.

3 Usability-Fragebögen

Viele Aspekte der Usability sind am besten zu erheben, indem man Benutzer befragt (Nielsen 1997). Neben der Erhebung objektiver Daten aus Usability-Tests können standardisierte Usability-Fragebögen eingesetzt werden, um subjektive Einstellungen und Erfahrungen von Usern zu messen. Zwischen objektiven Messungen (wie Aufgabenperformanz und Fehlerraten) und subjektiver Einschätzung der Zufriedenheit mit dem System besteht ein starker positiver Zusammenhang (Nielsen and Levy 1994). Usability-Fragebögen eignen sich im Besonderen dafür, jene Aspekte eines Softwareproduktes, die hinsichtlich der Usability problematisch sind, aufzudecken und verschiedene Systeme miteinander zu vergleichen. Da es schwierig ist, auf Grund von Rating Skalen im Detail zu verstehen, warum ein Produkt weniger gebrauchstauglich ist, und aufgrund des Interesses, konkrete Verbesserungsmöglichkeiten zu finden, inkludieren viele vorhandene Usability-Fragebögen neben Ratingskalen auch offene Antwortfelder für die Beschreibung konkreter Usability-Probleme.

Es existieren zahlreiche Usability-Fragebögen; im englischen Sprachraum gelten die Fragebögen Questionnaire for User Interface Satisfaction - QUIS (Chin, Diehl et al. 1988), Software Usability Measurement Inventory - SUMI (Kirakowski and Corbett 1993; Kirakowski 1996), die Items zu „perceived ease of use“ (Davis 1989) und der Post Study Usability Questionnaire (PSSUQ) (Lewis 2002) als besonders einflussreich. Im deutschen Sprachraum haben sich Übersetzungen dieser Usability-Fragebögen kaum etabliert, vielmehr wurden eigene Fragebögen entwickelt und publiziert. Im Folgenden wird genauer auf die deutschsprachigen Usability-Fragebögen Isonorm und Isometrics, die sich beide an der EN ISO 9241-110 orientieren, eingegangen. ErgoNorm (Dzida, Hofmann et al. 2000) ist ebenfalls ein deutscher Fragebogen, in dem die EN ISO 9241-110 operationalisiert wurde. Dieser ist im Vergleich zu Isonorm und Isometrics, welche mit Hilfe von Ratingskalen Usability zu quantifizieren beabsichtigen, auf Grund der Fragengestaltung und der offenen Antwortfelder stärker darauf ausgerichtet spezifische Nutzungsprobleme zu identifizieren. Auf weitere deutsche Usability-Fragebögen sei bloß hingewiesen, z.B. auf EU-CON II (Stary, Riesenecker-Caba et al. 1997) oder auf den Fragebogen zur Bewertung von Software von IfADo (Institut für Arbeitsphysiologie (IfADo) 1995) .

3.3 Isonorm 9241/10

Der Usability-Fragebogen Isonorm 9241/10 wurde 1993 veröffentlicht (Prümper and Anft 1993). Die Skalen des Fragebogens stellen eine Operationalisierung der 7 Kriterien der gleichnamigen EN ISO 9241-10 (Europäisches Komitee für Normung 1995) dar. Der Fragebogen wurde entwickelt, um die Normkonformität nach EN ISO 9241-10 überprüfen und um Prototypen bei iterativer Softwareentwicklung evaluieren zu können (Prümper 1997). Der Fragebogen ist für das Ausfüllen durch Benutzer vorgesehen, und die Durchführung dauert ca. 10 Minuten. Insgesamt besitzt der Fragebogen 35 Items, wobei zu jeder der 7

Skalen fünf Items gehören, die aus bipolaren Aussagen bestehen. Das Antwortformat (siehe Abb. 1) stellt eine 7-stufigen Ratingskala dar, mit der von sehr negativ ("---") bis sehr positiv ("+++") die bipolaren Aussagen beurteilt werden.

Aufgabenangemessenheit

Unterstützt die Software die Erledigung Ihrer Arbeitsaufgaben, ohne Sie als Benutzer unnötig zu belasten?

Die Software...	---	--	-	-/+	+	++	+++	
ist kompliziert zu bedienen.	○	○	○	○	○	○	○	ist unkompliziert zu bedienen.

Abb. 1 Beispielitem aus dem Fragebogen Isonorm 9241/10

Fig. 1 Example item from the Isonorm 9241/10 questionnaire

3.4 Isometrics

In dem Fragebogen Isometrics sind ebenfalls die 7 Gestaltungsgrundsätze von der EN ISO 9241-10 operationalisiert (Hamborg 2002). Die Gestaltungsgrundsätze werden mit jeweils 7 bis 12 Items erhoben. Insgesamt besteht der Fragebogen aus 75 Items, die auf einer 5-stufigen Ratingskala beantwortet werden. Es besteht auch die Möglichkeit „Keine Angabe“ anzukreuzen, wie in Abb. 2 gesehen werden kann. Es existiert eine Kurzversion des Fragebogen Isometrics^{S} (short) für summative und eine Langversion Isometrics^{L} (long) für formative Evaluationen. Diese beiden Versionen unterscheiden sich darin, dass in der Langversion zusätzlich zu jedem Item ein Wichtigkeitsrating gegeben werden kann. Außerdem sollen die Probanden in der Langform ihre Antworten begründen, indem sie Beispiele für Mängel notieren.

		stimmt nicht	stimmt wenig	stimmt mittelmäßig	stimmt ziemlich	stimmt sehr	
Index	**Aufgabenangemessenheit**	**1**	**2**	**3**	**4**	**5**	Keine Angabe
A.1	Die Software zwingt mich, überflüssige Arbeitsschritte durchzuführen.						

Abb. 2 Beispielitem aus dem Fragebogen Isometrics

Fig. 2 Example item from the Isometrics questionnaire

3.5 Vergleich der Usability-Fragebögen anhand von Qualitätskriterien

Zur Beurteilung von Fragebögen können Qualitätskriterien wie Objektivität, Reliabilität und Validität (Rost 2004) herangezogen werden, welche in diesem Kontext oft auch als „Testgütekriterien“ bezeichnet werden und welche im Folgenden im Detail vorgestellt werden. Darüber hinaus können Nebengütekriterien wie Normierung, Ökonomie und Nützlichkeit eine wertvolle Entscheidungshilfe bei der Auswahl eines Instruments bieten. Manche dieser Kriterien sind auf Basis der Fragebögen selbst zu beurteilen und benötigen keine weitere empirische Untersuchung. Diese werden daher gleich im Anschluss herangezogen, um die Qualität der deutschsprachigen Usability-Fragebögen Isonorm und Isometrics zu diskutieren. Die Autoren dieser Usability-Fragebögen haben auch in entsprechenden Artikeln Ergebnisse zu Untersuchungen der Testgütekriterien publiziert, welche ebenfalls beschrieben werden (Willumeit, Gediga et al. 1995; Hamborg, Willumeit et al. 1996; Prümper 1997; Gediga, Hamborg et al. 1999; Richter 1999). Da sich bisherige Untersuchungen jedoch nicht einem Vergleich der Usability-Fragebögen widmeten, sondern

jeweils nur einen Usability-Fragebogen betrachteten, können erhobene Werte nur annähernd verglichen werden.

3.5.1 Objektivität

Das Gütekriterium der Objektivität wird durch das Faktum bewertet, „in welchem Ausmaß die Testergebnisse vom Testanwender unabhängig sind“ (Bortz and Döring 2002). Durch die standardisierte, schriftliche Gestaltung der Fragebögen mit Rating-Skalen sowie klaren Auswertungsvorschriften für die Berechnung der einzelnen Summenscores ist das Kriterium der Objektivität sowohl beim Usability-Fragebogen Isonorm als auch Isometrics erfüllt. Die Testdurchführung ist immer vergleichbar, verschiedene Auswerter würden anhand eines ausgefüllten Fragebogens zu gleichen Scores für die Usability-Bewertung gelangen und könnten die Testergebnisse anhand der Manuale ähnlich interpretieren.

3.5.2 Validität

Das Gütekriterium der Validität ist definiert als das „Ausmaß, in dem ein Test das misst, was er zu messen vorgibt“ (Bühner 2006), und lässt sich in Inhalts-, Kriteriums- und Konstruktvalidität unterteilen. Inhaltsvalidität ist gegeben, wenn das zu messende Konstrukt hinreichend genau und erschöpfend erfasst wird (Bortz and Döring 2002; Bühner 2006). Da das zu messende Konstrukt in der EN ISO 9241-10 genau beschrieben ist, beide Usability-Fragebögen in ihrem Aufbau der Norm folgen und pro Grundsatz je eine Skala mit einer Reihe Items anbieten, die augenscheinlich die Inhalte der Norm operationalisieren, kann die Inhaltsvalidität für beide Usability-Fragebögen in gleichem Ausmaß als gegeben angesehen werden. Im Gegensatz zu Inhaltsvalidität, für die kein Kennwert herangezogen werden kann und die daher letztlich immer auf einer subjektiven Einschätzung basiert, sind die Kriteriums- und Konstruktvalidität auch intersubjektiv statistisch überprüfbar.

Mit Kriteriumsvalidität bezeichnet man den Zusammenhang zwischen einem Fragebogenergebnis und einem Außenkriterium. Das wichtigste Aussenkriterium im Kontext der Usability-Messung ist das Faktum, ob die Benutzer mit Hilfe der Software vorgegebene Ziele erreichen können und welchen Aufwand (z.B. an Zeit) sie dafür benötigen. Da Usability-Messungen grundsätzlich auch dafür dienen, alternative Designs oder verschiedene Softwareprodukte miteinander zu vergleichen, sollte die Messung hinreichend genau sein, um auch bei geringen Usability-Unterschieden zwischen unterschiedlichen Produkten differenzieren zu können.

Konstruktvalidität liegt vor, wenn „aus dem zu messenden Zielkonstrukt Hypothesen ableitbar sind, die anhand der Testwerte bestätigt werden können“ (Bortz and Döring 2002). Oftmals wird dabei überprüft, ob das Ergebnis mit anderen Messungen konstruktverwandter Verfahren erwartungsgemäß hoch (konvergente Validität), mit konstruktfernen Kriterien hingegen niedrig korreliert (diskriminante Validität).

Beim Fragebogen Isometrics wurde bereits bei der Entwicklung Wert auf inhaltliche Validität durch theoretisch fundierte Auswahl der Items gelegt (Hamborg, Willumeit et al. 1996; Gediga, Hamborg et al. 1999). Zuerst wurden 651 Items aus anderen Usability-Fragebögen und Checklisten gesammelt. Dann wurden redundante oder zu kompliziert formulierte Items entfernt und von sechs Usability-Experten den 7 Grundsätzen der EN ISO 9241-10 zugeordnet (Interrater-Reliabilität: Kappa = 0.75). Für eine Validierung wurden weiters die Skalenmittelwerte verschiedener Programme miteinander verglichen, und die Bewertungsunterschiede entsprachen den Erwartungen (Hamborg, Willumeit et al. 1996; Gediga, Hamborg et al. 1999).

Für den Fragebogen Isonorm wurden die Konstruktvalidität sowie die innere und äußere kriterienbezogene Validität überprüft. Die Konstruktvalidität wurde getestet, indem man Beurteilungen jeweils zweier Versionen von MS-Word und MS-Excel verglich. Es zeigte sich, dass anhand des Fragebogens Isonorm Veränderungen zwischen verschiedenen Versionen einer Software gemessen werden können. In den Untersuchungen zur inneren kriterienbezogenen Validität wurden Bewertungen der Usability-Fragebögen Isonorm, QUIS (Chin, Diehl et al. 1988) und BBD (Spinas 1987) korreliert, die zeigten, dass der Fragebogen Isonorm zu vergleichbaren Ergebnissen führte. Die äußere kriterienbezogene Validität wurde mit Hilfe von Expertenbefragungen nach EVADIS 2 bestätigt.

3.5.3 Reliabilität

Reliabilität bzw. Zuverlässigkeit gibt den „Grad der Genauigkeit, mit dem das geprüfte Merkmal gemessen wird“ an (Bortz and Döring 2002). Bei der Retest-Reliabilität, überprüft man, ob ein wiederholter Test zum selben Messergebnis führt, bei der Paralleltestreliabilität, ob eine zweite Testversion, die dasselbe operationalisiert, zum selben Ergebnis gelangt (Bühner 2006). Die Reliabilität kann auch ohne zusätzliche Messung durch die Testhalbierungs-Methode oder die Bestimmung der internen Konsistenz gemessen werden. Der am häufigsten verwendete Test zur Überprüfung der internen Konsistenz ist Cronbach’s α. Um zu gewährleisten, dass die Items ein Merkmal unidimensional messen und in einer Skala zusammengefasst werden können, sollte Cronbach’s α größer als 0.8 sein (Nunnally and Bernstein 1994).

Die Konsistenzanalyse bescheinigte dem Fragebogen Isonorm eine gute innere Konsistenz der einzelnen Skalen (Cronbach’s alpha von 0.81 bis 0.89). Hinsichtlich der Reliabilitätsbestimmung weist der Fragebogen Isonorm eine hohe Retest-Reliabilität auf (von $r_{t1/t2} = 0.59$ bis zu $r_{t1/t2} = 0.68$).

Auch die Reliabilitätswerte des Fragebogens Isometrics werden von den Autoren hinsichtlich der statistischen Qualität der Skalen als zumindest zufrieden stellend beschrieben (Willumeit, Gediga et al. 1995; Hamborg, Willumeit et al. 1996).

3.5.4 Nebengütekriterien

In Bezug auf die Nebengütekrierien liegen für alle drei Usability-Fragebögen **Normen** für verschiedene Softwareprodukte vor, die hilfreich sein können, erhobene Evaluierungsdaten einzuordnen. Für den Fragebogen Isonorm liegen allgemeine Normwerte für Softwaresysteme vor, die auf der Beurteilung von 41 Programmen basieren (Prümper 1997), wobei die Skala Individualisierbarkeit generell die niedrigste, die Skala Steuerbarkeit die höchste Bewertung erhielt. Auch zu Isometrics liegen Normtabellen zu mehreren bekannten Softwareprogrammen vor (z.B. MS Word, SAP/Term).

Da beide Usability-Fragebögen dieselbe Norm operationalisieren, und dieselben spezifischen Usability-Aspekte messen, ist das Ausmaß ihrer **Nützlichkeit** für den jeweiligen Evaluierungskontext vergleichbar. Der Fragebogen Isonorm ist z.B. neben der Papierversion auch als Onlineversion erhältlich, welche dieselben Ergebnisse liefert (Richter 1999). Der Fragebogen Isometrics liegt in unterschiedlichen Versionen für Gruppen- und Einzeltestung, sowie für formative und summative Evaluation vor (Hamborg, Gediga et al. 1999). Hinsichtlich der Nützlichkeit der Langversion Isometrics^{L} zeigte sich im Vergleich mit den Evaluationsmethoden Lautes Denken, Videokonfrontation und Heuristische Evaluation, dass in Isometrics^{L} 3-5 mal so viele Anmerkungen über Probleme der Software als bei den anderen gemacht wurden (Hamborg 2002).

Für die Beurteilung der **Ökonomie** können die Angaben aus den Manualen der beiden Usability-Fragebögen herangezogen werden: Wie viele Items beinhaltet der Fragebogen? Wie lange dauert die Durchführung und Auswertung? Bei diesem Punkt ist Isometrics mit 75 Items deutlich umfangreicher als Isonorm mit 35 Items.

4 Fragestellungen

Wie aus der vorangegangenen Darstellung bereits vorliegender Studien zu den Usability-Fragebögen Isometrics und Isonorm ersichtlich wurde, erfüllen beide Fragebögen die testtheoretischen Qualitätsanforderungen. Offen bleibt jedoch die Frage, ob einem der Usability-Fragebögen für empirische Untersuchungen der Vorzug gegeben werden sollte bzw. inwiefern Usability-Messungen der Fragebögen vergleichbar sind. Um die Fragebögen Isonorm und Isometrics näher zu vergleichen, werden die Hauptgütekriterien Reliabilität und Validität sowie Nebengütekriterien wie die Präferenz der Benutzer für einen der beiden Fragebögen untersucht.
Hinsichtlich des Validität der Usability-Fragebögen Isometrics und Isonorm ist es eine zentrale Frage, ob Software in beiden Fragebögen in ähnlicher Weise beurteilt wird, d.h. ob die beiden Fragebögen vergleichbare Evaluationsresultate liefern. Außerdem ist es von Interesse, ob die Messungen beider Usability-Fragebögen mit Aussenkriterien wie Bearbeitungszeit und Aufgabenlösung in einem Usability-Test zusammenhängen.
Auch in Bezug auf Nebengütekriterien ergeben sich relevante offene Fragestellungen – insbesonders ob Benutzer das Ausfüllen der Usability-Fragebögen subjektiv unterschiedlich erleben und wie lange und intensiv Nutzer mit der Software vertraut sein müssen, um die Fragebögen ausfüllen zu können.

5 Methodik

Für die Beantwortung der Fragestellungen wurde ein experimentelles Zwei-Versuchsgruppendesign gewählt: Teilnehmer der Versuchsgruppe 1 lernten das Statistikprogramm SPSS, Teilnehmer der Versuchsgruppe 2 das Statistikprogramm STATISTICA kennen und beurteilten im Anschluss die Usability dieses Programms anhand der Fragebögen Isonorm und Isometrics (siehe Abb. 3). Bei den gewählten Standard-Softwarepaketen, die sich beide auf die Analyse sozialwissenschaftlicher Daten konzentrieren, war gewährleistet, dass im Usability-Test jeweils dieselben Aufgaben gestellt werden konnten, d.h. dieselben funktionalen Anforderungen waren mittels zwei verschiedener User-Interfaces realisiert. Die Versuchsgruppen waren je nach Reihenfolge der Vorlage der Fragebögen Isonorm und Isometrics in Subversuchsgruppen eingeteilt.

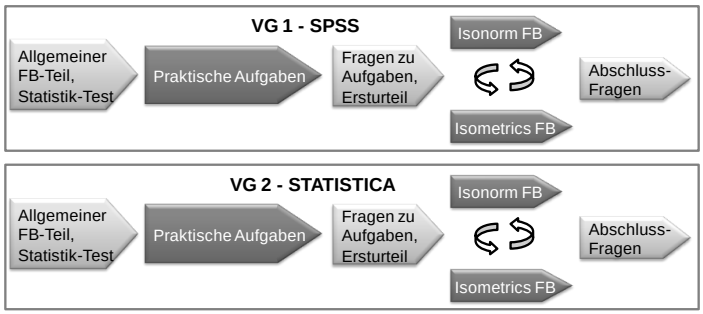
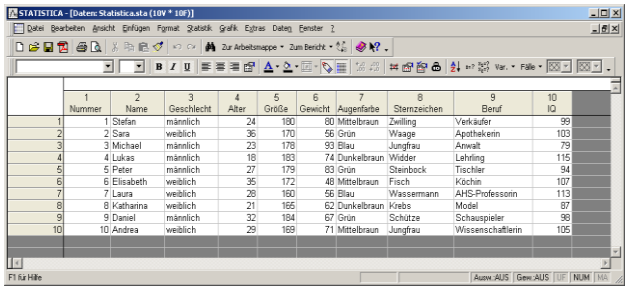


Abb. 3 Überblick über das Versuchsdesign

Fig. 3 Overview of the research design

5.1 Usability-Test

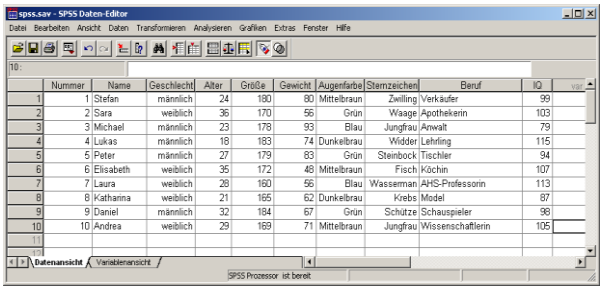
Der Usability-Test wurde auf zwei Notebooks durchgeführt um eine einheitliche Testumgebung zu garantieren. Anhand von 14 Aufgaben mit derselben Datendatei mit 10 Datensätzen und je 10 Variablen lernten die Versuchspersonen die beiden Programme SPSS bzw. STATISTICA kennen. Die Abb. 4 und 5 zeigen Screenshots der Programmansicht am Anfang des Usability-Tests, nachdem das Datenfile durch die Versuchspersonen geöffnet wurde.



	1 Nummer	2 Name	3 Geschlecht	4 Alter	5 Größe	6 Gewicht	7 Augenfarbe	8 Sternzeichen	9 Beruf	10 IQ
1	1	Stefan	männlich	24	180	80	Mittelbraun	Zwilling	Verkäufer	99
2	2	Sara	weiblich	36	170	56	Grün	Waage	Apothekerin	103
3	3	Michael	männlich	23	178	93	Blau	Jungfrau	Anwalt	79
4	4	Lukas	männlich	18	183	74	Dunkelbraun	Widder	Lehrling	115
5	5	Peter	männlich	27	179	83	Grün	Steinbock	Tischler	94
6	6	Elisabeth	weiblich	35	172	48	Mittelbraun	Fisch	Köchin	107
7	7	Laura	weiblich	28	160	56	Blau	Wassermann	AHS-Professorin	113
8	8	Katharina	weiblich	21	165	62	Dunkelbraun	Krebs	Model	87
9	9	Daniel	männlich	32	184	67	Grün	Schütze	Schauspieler	98
10	10	Andrea	weiblich	29	169	71	Mittelbraun	Jungfrau	Wissenschaftlerin	105

Abb. 4 STATISTICA Arbeitsoberfläche mit Datenfile der Untersuchung

Fig. 4 STATISTICA user interface with the data file from the experiment



	Nummer	Name	Geschlecht	Alter	Größe	Gewicht	Augenfarbe	Sternzeichen	Beruf	IQ
1	1	Stefan	männlich	24	180	80	Mittelbraun	Zwilling	Verkäufer	99
2	2	Sara	weiblich	36	170	56	Grün	Waage	Apothekerin	103
3	3	Michael	männlich	23	178	93	Blau	Jungfrau	Anwalt	79
4	4	Lukas	männlich	18	183	74	Dunkelbraun	Widder	Lehrling	115
5	5	Peter	männlich	27	179	83	Grün	Steinbock	Tischler	94
6	6	Elisabeth	weiblich	35	172	48	Mittelbraun	Fisch	Köchin	107
7	7	Laura	weiblich	28	160	56	Blau	Wasserman	AHS-Professorin	113
8	8	Katharina	weiblich	21	165	62	Dunkelbraun	Krebs	Model	87
9	9	Daniel	männlich	32	184	67	Grün	Schütze	Schauspieler	98
10	10	Andrea	weiblich	29	169	71	Mittelbraun	Jungfrau	Wissenschaftlerin	105

Abb. 5 SPSS Arbeitsoberfläche mit Datenfile der Untersuchung

Fig. 5 SPSS user interface with the data file from the experiment

Bei den 14 Aufgaben wurde darauf geachtet, dass die Versuchspersonen ein möglichst umfassendes Bild der Software erlangen und sowohl Datenmanipulation, deskriptive Statistiken, statistische Analysemodule, Graphikerstellung und die Hilfefunktionen kennen lernen. Auch wurden absichtlich drei Eingabefehler in der Angabe eingebaut, damit die Versuchspersonen die Fehlertoleranz des Programms testen und Fehlermeldungen kennen lernen konnten. Die Aufgaben im praktischen Usability-Test wurden in drei verschiedenen Schwierigkeitsgraden gestellt. Am Anfang gab es Aufgaben mit vollständiger Anleitung, in der jeder Mausklick beschrieben wurde, z.B. eine neue Variable definieren oder statistische Analysen wie t-Test oder Korrelation durchführen. Dann gab es Aufgaben, in denen zuvor erklärte Verfahren bzw. Graphiken ohne Anleitung selbständig mit anderen Variablen ausgeführt werden sollten. Abschließend sollten völlig neue Prozeduren ohne Erklärung gelöst werden, z.B. Variablen sortieren, kopieren oder löschen sowie Graphiken erstellen oder Verfahren wie die Regressionsanalyse anwenden und die Hilfefunktionen verwenden.

Es wurde versucht, den Wortlaut der Anleitungen in der SPSS- sowie in der STATISTICA-Version soweit wie möglich gleich zu halten. Die Abb. 6 zeigt ein Beispiel einer Aufgabenanleitung für das Erstellen von Histogrammen. Nach Abschluss der Aufgaben wurden die Versuchspersonen gebeten, die veränderte Datendatei und die Ausgabedatei zu speichern.

Aufgabe 5: Histogramm
• Wählen Sie aus dem Menü *Grafik -> Histogramme.* • Klicken Sie links oben auf *Variablen.* • Klicken sie auf *„Gewicht“* um die Variable Gewicht auszuwählen. • Klicken Sie auf *OK.* • Klicken Sie rechts unten auf *OK.*

Abb. 6 Beispiel einer Aufgabenanleitung

Fig. 6 Example of task instructions for an experimental task

5.2 Einsatz von Isonorm und Isometrics

Für die entsprechenden Subversuchsgruppen gab es eine Fragebogenversion, in der zuerst der Fragebogen Isonorm, und eine andere, in der zuerst Isometrics zu beantworten war. Die Usability-Fragebögen wurden im Original übernommen, mit der Ausnahme, dass die Antwortmöglichkeit „Keine Angabe“ beim Isometrics Fragebogen weggelassen wurde, um Vergleichbarkeit mit dem Isonorm Fragebogen zu gewährleisten. Ergänzt wurden alle vorgegebenen Items der beiden Fragebögen um jeweils ein zusätzliches Item, in dem auch der eigene Kenntnisstand zum Beurteilen des Items eingeschätzt werden sollte.

Für den Vergleich von Isometrics und Isonorm wurden drei Fragen am Ende des Fragebogens gestellt. Die Versuchspersonen wurden gebeten, einen direkten Vergleich hinsichtlich folgender Dimensionen zu machen: welchen Fragebogen sie angenehmer zu beantworten gefunden haben, bei welchem sie das Gefühl hatten, ihre Meinung besser ausdrücken zu können und welchen Fragebogen sie unabhängig von der unterschiedlichen Anzahl von Items schneller beantworten konnten.

5.3 Stichprobe

Die Stichprobe bestand aus 50 Versuchspersonen, davon je 50% weiblich bzw. männlich. Die Personen waren von 21 bis 34 Jahre alt und hatten zumindest Matura/Abitur als höchste abgeschlossene Schulbildung. Um sicherzustellen, dass es nicht zu Verzerrungen auf Grund unterschiedlicher Vorerfahrung mit ähnlichen Softwareprodukten kommt, wurde als Ausschlusskriterium für die Studienteilnahme festgelegt, dass die Teilnehmer weder mit SPSS, noch mit STATISTICA oder vergleichbaren Statistikprogrammen bereits gearbeitet hatten. Ein Großteil der Versuchspersonen (88%) gab an, Grundlagen der Statistik schon in der Schule und/oder an der Universität kennen gelernt zu haben. Die Versuchspersonen hatten den Eindruck, dass die Anleitung leicht verständlich und die Aufgaben leicht lösbar waren; ihr Statistikwissen hielten sie im Wesentlichen ausreichend, um den Inhalt der Aufgaben zu verstehen. Insofern war gewährleistet, dass die Versuchspersonen den Usability-Test im vorgesehenen Sinne durchführen konnten.

6 Ergebnisse

Die Gesamt-Durchführungsdauer des Experiments lag bei ca. 50 Minuten, wobei im Durchschnitt 26 Minuten für die Interaktion mit der Software und die restliche Zeit für das Ausfüllen der Usability-Fragebögen verwendet wurde.

6.1 Validität

6.1.1 Kriteriumsvalidität: Aussenkriterium „Unterschiedliche Softwareprodukte“

Zur Überprüfung, ob die beiden Usability-Fragebögen zwischen den beiden Statistik-Standardsoftwarepaket differenzieren können bzw. ob der Einsatz der beiden Fragebögen dasselbe Ergebnis hinsichtlich der Messung der Usability liefert, wurden die jeweiligen Bewertungen miteinander verglichen. Dazu wurde eine zweifaktorielle multivariate Varianzanalyse mit Messwiederholung ($df_{Hypothese} = 6$, $df_{Fehler} = 39$) gerechnet, mit den Usability-Fragebögen als Innersubjektfaktor und der Versuchsbedingung der unterschiedlichen Statistikprogramme als Zwischensubjektfaktor. Der Innersubjektfaktor

„Usability-Fragebogen“ (Isonorm vs. Isometrics) hatte keinen Einfluss auf die Beurteilung der Usability ($F=1.21$, $p=0.32$, $\eta^2=0.16$). Daraus kann abgeleitet werden, dass die beiden Usability-Fragebögen dasselbe Messergebnis lieferten. Weiters zeigt sich, dass der Zwischensubjektfaktor „Statistikprogramm“ (SPSS vs. STATISTICA) einen signifikanten Einfluss auf die Bewertung der Usability hat ($F=3.40$, $p=0.009$, $\eta^2=0.34$), d.h. die Usability der beiden Softwareprodukte wurde unterschiedlich bewertet (siehe Abb. 7 und Abb. 8). Beide Usability-Fragebögen unterstützten das Resultat, dass SPSS eine höhere Steuerbarkeit als STATISTICA aufwies. In der Dimension Steuerbarkeit erreichte es in beiden Usability-Fragebögen bessere Werte, in der Dimension Selbstbeschreibungsfähigkeit erhielt es nur im Fragebogen Isometrics eine bessere Beurteilung.
Auch wenn man nur Beurteilungen in dem jeweils erst gereihten Usability-Fragebogen miteinander vergleicht, die noch nicht durch eine vorhergehende Beurteilung beeinflusst waren, kommt man zu demselben Ergebnis. Dass die hohe Übereinstimmung der Messergebnisse aufgrund der sequentiellen Beurteilung in beiden Fragebögen, zustande gekommen ist, kann somit ausgeschlossen werden.
Für die Interpretation der Ergebnisse wurden zusätzlich U-Tests für Unterschiede auf Itemebene gerechnet. Dabei konnten teilweise ähnliche Aussagen anhand der beiden Usability-Fragebögen getroffen werden, sofern inhaltlich vergleichbare Itemformulierungen in einer Skala vorhanden waren. Hinsichtlich Steuerbarkeit zeigte sich zum Beispiel, dass die Versuchspersonen in SPSS stärker das Gefühl hatten, an jedem Punkt ihre Arbeit ohne Verluste unterbrechen und leichter zwischen Menüs und Masken wechseln zu können, als dies bei STATISTICA der Fall war. Die Probanden hatten in SPSS stärker als in STATISTICA das Gefühl, dass es jederzeit möglich wäre, eine Befehlseingabe abzubrechen. Die Ergebnisse für die Items der Skala Selbstbeschreibungsfähigkeit, veranschaulichten, dass die Versuchspersonen in SPSS eher meinten, bei Bedarf Erläuterungen und Informationen zu Eingabefeldern abrufen zu können als in STATISTICA. Außerdem scheint es in SPSS leichter zu erkennen zu sein, wenn Befehle nicht zur Verfügung stehen, und die Software scheint eher Informationen über zulässige Eingaben zu liefern als in STATISTICA.

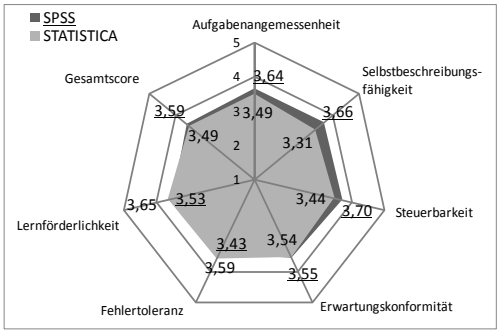


Abb. 7 Beurteilung der Usability im Fragebogen Isometrics (Mittelwerte)

Fig. 7 Usability assessment with the Isometrics questionnaire (means)

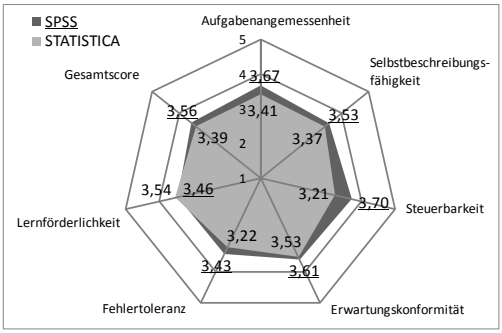


Abb. 8 Beurteilung der Usability im Fragebogen Isonorm (Mittelwerte)

Fig. 8 Usability assessment with the Isonorm questionnaire (means)

6.1.2 Kriteriumsvalidität: Aussenkriterien „Bearbeitungszeit und Anzahl gelöster praktischer Aufgaben“

Zusammenhänge mit objektiven Messzahlen aus Usability-Tests (wie Bearbeitungszeit und Vollständigkeit der Aufgabenlösung) können herangezogen werden um die Validität einer Fragebogenmessung zu überprüfen.

In dem erhobenen Datensatz zeigten sich erwartungsgemäß ausschließlich negative Korrelationen zwischen der Bearbeitungszeit und der Bewertung der Usability, wie in Tab. 1 zu sehen ist. Sämtliche Korrelationen mit Ausnahme von Aufgabenangemessenheit und Lernförderlichkeit im Fragebogen Isonorm, und der Erwartungskonformität im Fragebogen Isometrics, sowie der Fehlertoleranz in beiden Fragebögen, sind mindestens auf einem Alpha Niveau von 0.05 signifikant. Dieses Resultat bedeutet, dass die Versuchspersonen die Usability der Software umso schlechter einschätzten, je länger sie sich mit den praktischen Aufgaben auseinander setzten, um diese zu lösen. Je nach Dimension variiert es, ob die Korrelationen mit der Messung in Isonorm bzw. Isometrics jeweils höher ausfallen, was zum Teil eventuell dadurch beeinflusst sein könnte, dass die Items einen unterschiedlich starken Bezug zur Aufgabenbearbeitung mit der Software haben.

Zwischen der Anzahl gelöster praktischer Aufgaben und der Beurteilung der Usability konnten keine Zusammenhänge festgestellt werden. Da dieses Ergebnis jedoch für die Messungen in beiden Fragenbögen zutraf, lässt sich dieses wohl auf die Art der Aufgaben zurückführen, die großteils mit ausführlichen Erklärungen zu bearbeiten waren und somit nur zu geringer Varianz in der Vollständigkeit der Aufgabenlösung führten.

Tab. 1 Pearson Korrelation zwischen Bearbeitungszeit, Anzahl der gelösten praktischen Aufgaben und Beurteilung der Usability-Dimensionen (n = 46-48) [In den Tabellen ist eine hochsignifikante Korrelation mit p<0.001 jeweils durch drei Sterne (***), mit p<0.01 durch zwei Sterne (**) und eine signifikante Korrelation mit p < 0.05 durch einen Stern (*) markiert, ein Trend ist mit (+) markiert.]

Tab. 1 Pearson correlation between execution time, number of tasks solved correctly and assessment of usability criteria (n = 46-48)

Fragebogen	Dimension	Bearbeitungszeit	Anzahl der gelösten praktischen Aufgaben
Isonorm	Aufgabenangemessenheit	-0.21	0.11
	Selbstbeschreibungsfähigkeit	**-0.37****	-0.07
	Steuerbarkeit	**-0.36****	0.23
	Erwartungskonformität	**-0.37****	0.12
	Fehlertoleranz	-0.19	0.10
	Lernförderlichkeit	-0.24	0.04
	Gesamtscore	**-0.36****	0.10
Isometrics	Aufgabenangemessenheit	**-0.43****	0.17
	Selbstbeschreibungsfähigkeit	**-0.48*****	0.02
	Steuerbarkeit	**-0.45*****	0.04
	Erwartungskonformität	**-0.27**$^{+}$	0.17
	Fehlertoleranz	-0.14	-0.15
	Lernförderlichkeit	**-0.47*****	0.26
	Gesamtscore	**-0.46*****	0.13

6.1.3 Konstruktvalidität: Zusammenhang zwischen den Usability-Fragebögen

Hinsichtlich des Zusammenhangs der Einschätzungen in den beiden Usability-Fragebögen, wurden Korrelationen für jede einzelne der 6 Skalen sowie für den Gesamtscore berechnet. Wie Tab. 2 zeigt, sind sämtliche Korrelationen positiv, die meisten sogar mittel bis hoch. Versuchspersonen schätzen somit die Dimensionen der Usability in beiden Fragebögen in ähnlicher Weise ein.

Tab. 2 Pearson Korrelationen zwischen den Beurteilungen im Isonorm und Isometrics Fragebogen (n = 48-49)

Tab. 2 Pearson correlation between usability assessments in Isonorm and Isometrics (n = 48-49)

Dimension	r
Aufgabenangemessenheit	**0.45*****
Selbstbeschreibungsfähigkeit	**0.51*****
Steuerbarkeit	**0.37****
Erwartungskonformität	**0.29****
Fehlertoleranz	**0.24**$^{+}$
Lernförderlichkeit	**0.66*****

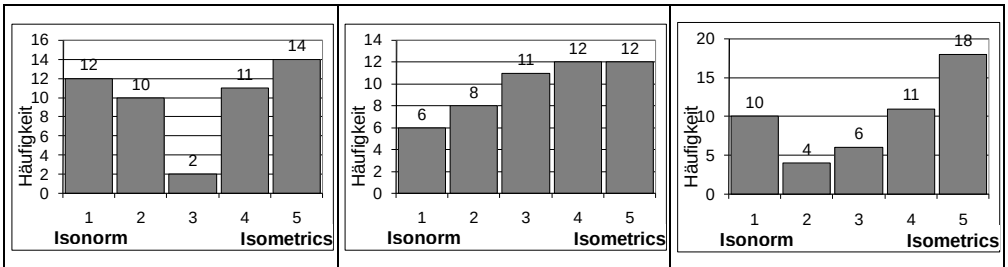
6.2 Reliabilität

Reliabilitätsanalysen zur Überprüfung der Frage, ob die Items einer Skala dieselbe Dimension messen, wurden für die Skalen der Fragebögen Isonorm und Isometrics jeweils getrennt für die Beurteilung der beiden Softwareprodukte gerechnet. Die Reliabilitäten können als durchwegs zufriedenstellend bewertet werden (Cronbach's Alpha von 0.54 - 0.91 bei allen eingesetzten Skalen). Zusätzliche Kennwerte zeigten, dass durch Ausschluss einzelner Items (z.B. „Das System lässt sich nur in einer starr vorgegebenen Weise bedienen") des Fragebogens Isometrics die Reliabilität der dazugehörigen Skalen etwas steigen würde. Dieses Ergebnis ist wahrscheinlich dadurch begründet, dass diese Items im Vergleich zu den anderen negativ formuliert waren.

6.3 Weitere Gütekriterien

6.3.4 Fragebogenbeurteilung der Versuchspersonen

Die Versuchspersonen beurteilten, ob sie einen der Usability-Fragebögen angenehmer bzw. schneller zum Beantworten empfunden haben, in drei abschließenden, vergleichenden Fragen, deren Ergebnisse in Abb. 9 zu sehen sind. Es scheint, dass wahrscheinlich auf Grund persönlicher Vorlieben für Antwortformate, je ein Teil entweder Isonorm oder Isometrics bevorzugt hat. Bei der Frage, ob es in einem der Usability-Fragebögen einfacher sei, seine Meinung besser auszudrücken als im anderen, hat es auch einige (11) indifferente Antworten gegeben. Die Frage, ob nun einer der beiden Usability-Fragebögen bei den Versuchspersonen „beliebter" war, kann nicht eindeutig beantwortet werden. Die Mittelwerte zeigen eine leichte Präferenz für den Fragebogen Isonorm, die Antwortverteilung lässt aber auch erkennen, dass es Versuchspersonen gegeben hat, denen der Isometrics Fragebogen mehr „liegt".



Welchen Fragebogen haben Sie angenehmer zum Beantworten gefunden?	Bei welchem Fragebogen hatten Sie das Gefühl ihre Meinung besser ausdrücken zu können?	Welchen Fragebogen konnten Sie schneller beantworten (unabhängig von unterschiedlicher Itemanzahl)?
Median = 4, MW = 3.10, SD = 1.61	Median = 3, MW = 3.33, SD = 1.34	Median = 4, MW = 3.47, SD = 1.56

Abb. 9 Beurteilung der Fragebögen Isonorm und Isometrics (n = 49)

Fig. 9 User assessment of the Isonorm and Isometrics questionnaires (n = 49)

6.3.5 Notwendiger Kenntnisstand für die Beantwortung der Items der Usability-Fragebögen

Zur Prüfung der Fragestellung, ob für das Ausfüllen der Usability-Fragebögen ein unterschiedlicher Kenntnisstand der Benutzer notwendig ist, ist eine multivariate Varianzanalyse mit Messwiederholung gerechnet worden. Es wurde deutlich, dass der Kenntnisstand für die Beantwortung der Items in den Usability-Fragebögen ($F_{df\_Hypothese=6,df\_Fehler=37}=2.87$, $p=0.02$) unterschiedlich eingeschätzt wurde. Dies läßt sich darauf zurückführen, dass der vorhandene Kenntnisstand bei der Skala Erwartungskonformität im Fragebogen Isonorm (MW = 3.20, SD = 0.99) höher als in Isometrics (MW = 2.86, SD = 1.08) eingeschätzt wurde. Versuchspersonen gaben bei Items, wie z.B. „Die Ausführung einer Funktion führt immer zu dem erwarteten Ergebnis“ (MW = 2.58, SD = 1.09) und „Die Bearbeitungszeiten der Software sind für mich gut abschätzbar“ (MW = 2.71, SD = 1.11) im Fragebogen Isometrics einen niedrigeren Kenntnisstand an als bei Items aus dem Fragebogen Isonorm, wie z.B. „Die Software lässt einen nicht im Unklaren darüber, ob eine Eingabe erfolgreich war oder nicht“ (MW = 3.34, SD = 0.98). Das Ergebnis ist insofern nicht verwunderlich, als das Abschätzen von Bearbeitungszeiten, oder der Tatsache ob ein erwartetes Ergebnis erreicht wurde, eher von fortgeschritteneren Benutzern als von Anfängern gemacht werden kann. Die entsprechenden Items des Isometrics-Fragebogens scheinen daher eher für eine Usability-Testung mit erfahrenen Benutzern geeignet zu sein.

7 Zusammenfassung und Ausblick

Diese Studie widmete sich dem Vergleich zweier weitverbreiteter, deutschsprachiger Usability-Fragebögen – Isonorm und Isometrics –, die beide die EN ISO 9241-110 operationalisieren. Zusammenfassend kann gesagt werden, dass die beiden untersuchten Usability-Fragebögen grundsätzlich dieselben Ergebnisse für Bewertung und Vergleich von Softwareprodukten liefern und ihre testtheoretische Qualität vergleichbar gut ist.

Die Studie zeigte ein hohes Maß an Übereinstimmung in der Messung mittels Isonorm und Isometrics und die Bewertungen im Isonorm- und Isometrics-Fragebogen korrelierten positiv miteinander. In beiden Usability-Fragebögen konnten zufriedenstellende Reliabilitäten festgestellt werden. Zusätzlich zeigte sich, dass für einige Items der Fragebogens Isometrics ein höherer Kenntnisstand des Systems notwendig ist als bei entsprechenden Items im Fragebogen Isonorm. Bei der Beurteilung der Usability-Fragebögen durch die Versuchspersonen wurde ebenfalls keiner eindeutig präferiert. Die Versuchspersonen wurden allerdings gebeten, dies unabhängig von der Itemanzahl zu beurteilen. Die Fragebogenbeurteilung deutet darauf hin, dass es sowohl Versuchspersonen gibt, die den Isonorm-Fragebogen bevorzugen, als auch solche, die den Isometrics Fragebogen als angenehmer und schneller zu beantworten finden. Da die Itemanzahl für die Zumutbarkeit jedoch ein relevanter Faktor ist, und das ökonomischere Verfahren Isonorm mit 35 Items weniger Zeit zum Ausfüllen als Isometrics mit 75 Items benötigt, sollte dies bei der Fragebogenauswahl auf jeden Fall mit bedacht werden, da prinzipiell beide in der Lage sind, dieselben Dimensionen zu messen.

Eine Möglichkeit für weitere Forschung stellt sicher die Tatsache dar, dass nur zwei Programme - und dies in einem spezifischen Anwendungsbereich - untersucht wurden.

Obwohl angenommen werden kann, dass das erhaltene Ergebnis über den spezifischen Kontext hinaus von Relevanz ist, wäre es durchwegs interessant, ob die Evaluierungen in den beiden Usability-Fragebögen in einer anderen Softwaresparte eventuell stärker abweichen würden.
Insbesondere wäre es auch von Interesse, mit einer größeren Stichprobe die faktorenanalytische Qualität der Usability-Fragebögen näher zu analysieren. In dem Datensatz der Studie zeigte sich als generelle Schwierigkeit der beiden Usability-Fragebögen, dass die Skalen im hohen Maße korrelieren, was ihre differenzielle Interpretation über den allgemeinen Usability-Faktor hinaus schwierig macht. Falls sich in weiteren testtheoretischen Untersuchungen zeigen würde, dass sich die einzelnen Usability-Prinzipien nicht unabhängig voneinander messen lassen, sondern es nur möglich ist, generelle Usability zu messen, so wäre das unter Umständen noch ökonomischer möglich als mit 35 bzw. 75 Items in den beiden Usability-Fragebögen.

Der vorliegende Beitrag rückt die Frage nach der Qualität von Messungen subjektiver Aspekte von Anwendungssystemen in den Vordergrund, die im wissenschaftlichen Diskurs immer stärker gefordert wird. Insofern möchte er einen Beitrag für die methodische Absicherung der Verwendung deutschsprachiger Usability-Fragebögen liefern.

Literatur

N. Bevan (2001). "International standards for HCI and usability." Int. J. Hum.-Comput. Stud. **55**(4): 533-552.

J. Bortz and N. Döring (2002). Forschungsmethoden und Evaluation. Berlin, Heidelberg, New York, Springer.

M. Bühner (2006). Einführung in die Test- und Fragebogenkonstruktion. München, Pearson.

J. P. Chin, V. A. Diehl, et al. (1988). Development of a Tool Measuring User Satisfaction of the Human-Computer Interface. SIGCHI '88 New York, ACM/SIGCHI.

F. D. Davis (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." MIS Quarterly **13**(3): 319-340.

W. H. Delone and E. R. McLean (2003). "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update." Journal of Management Information Systems **19**(4): 9-30.

W. Dzida, B. Hofmann, et al. (2000). Gebrauchstauglichkeit von Software. Ergonorm: Ein Verfahren zur Konformitätsprüfung von Software auf der Grundlage von DIN EN ISO 9241 Teile 10 und 11. Schriftenreihe der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. Dortmund, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.

Europäisches Komitee für Normung (1995). EN ISO 9241-10. Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten. Teil 10: Grundsätze der Dialoggestaltung.

Europäisches Komitee für Normung (1995). EN ISO 9241-11. Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten. Teil 11: Anforderungen an die Gebrauchstauglichkeit. Leitsätze.

Europäisches Komitee für Normung (2006). EN ISO 9241-110. Ergonomie der Mensch-System-Interaktion. Teil 110: Grundsätze der Dialoggestaltung.

G. Gediga and K.-C. Hamborg (2002). "Evaluation in der Software-Ergonomie: Methoden und Modelle im Software-Entwicklungsprozess." Zeitschrift für Psychologie **210**(1): 40-57.

G. Gediga, K.-C. Hamborg, et al. (1999). "The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10 supporting summative and formative evaluation of software systems." Behaviour and Information Technology **18**: 151-164.

K.-C. Hamborg (2002). Gestaltungsunterstützende Evaluation von Software: Zur Effektivität und Effizienz des IsoMetricsL Verfahrens. Mensch & Computer 2002: Vom interaktiven Werkzeug zu kooperativen Arbeits- und Lernwelten. M. Herczeg, W. Prinz and H. Oberquelle. Stuttgart, Teubner**:** 303-312.

K.-C. Hamborg, G. Gediga, et al. (1999). Softwareevaluation in Gruppen oder Einzelevaluation: Sehen zwei Augen mehr als vier? Softwareergonomie '99: Design von Informationswelten. U. Arend and K. Pitschke. Stuttgart, Teubner**:** 97-109.

K.-C. Hamborg, H. Willumeit, et al. (1996). Untersuchungen zu Itemformulierungen des IsoMetrics-Verfahrens. Osnabrücker Schriftenreihe Software-Ergonomie - 1. Osnabrück, Universität Osnabrück.

M. Herczeg (2005). Software-Ergonomie - Grundlagen der Mensch-Computer-Kommunikation. München, Wien, Oldenbourg Verlag.

K. Hornbæk (2006). "Current practice in measuring usability: Challenges to usability studies and research." International Journal of Human-Computer Studies **64**(2): 79-102.

J. Hüttner, H. Wandke, et al. (1995). Benutzerfreundliche Software- Psychologisches Wissen für die ergonomische Schnittstellengestaltung. Berlin, Bernd-Michael Paschke Verlag.

Institut für Arbeitsphysiologie (IfADo) (1995). Fragebogen zur Bewertung von Software, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA).

J. Kirakowski (1996). The Software Usability Measurement Inventory: background and usage. Usability Evaluation in Industry P. W. Jordan, B. Thomas, B. A. Weerdmeester and I. L. McClelland. London, Taylor & Francis**:** 169–177.

J. Kirakowski and M. Corbett (1993). "SUMI: The Software Usability Measurement Inventory." British Journal of Educational Technology **24**(3): 210-212.

J. R. Lewis (2002). "Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies." International Journal of Human-Computer Interaction **14**(3&4): 463–488.

J. Nielsen (1996). "Usability Metrics: Tracking Interface Improvements." IEEE Software **13**(6): 12-13.

J. Nielsen (1997). "Let's Ask the Users." IEEE Software **14**(3): 110-111.

J. Nielsen and J. Levy (1994). "Measuring usability: preference vs. performance." Communications of the ACM **37**(4): 66-75.

J. C. Nunnally and I. H. Bernstein (1994). Psychometric Theory. New York, McGraw-Hill.

J. Prümper (1997). Der Benutzungsfragebogen ISONORM 9241/10: Ergebnisse zur Reliabilität und Validität. Software-Ergonomie '97, Stuttgart, Reports of the German Chapter of the ACM.

J. Prümper and J. Anft (1993). Die Evaluation von Software auf Grundlage des Entwurfs zur internationalen Ergonomie-Norm ISO 9241 Teil 10 als Beitrag zur partizipativen Systemgestaltung - ein Fallbeispiel. Software-Ergonomie '93, Stuttgart, Teubner.

M. Rauterberg (1992). "Lässt sich die Gebrauchstauglichkeit interaktiver Software messen? Und wenn ja, wie?" Ergonomie & Informatik **16**: 3-18.

M. Richter (1999). Online Befragung als neues Instrument zur Beurteilung der Benutzerfreundlichkeit von Software. Current Internet science - trends, techniques, results. Aktuelle Online Forschung - Trends, Techniken, Ergebnisse. E.-D. Reips, B. Batanic, W. Bandillaet al. Zürich, Online Press.

J. Rost (2004). Lehrbuch Testtheorie-Testkonstruktion. Bern, Göttingen, Toronto, Seattle, Verlag Hans Huber.

P. Spinas (1987). Arbeitspsychologische Aspekte der Benuzerfreundlichkeit von Bildschirmsystemen. Zürich, ADAG.

C. Stary, T. Riesenecker-Caba, et al. (1997). EU-CON. ein Verfahren zur EU-konformen Software-ergonomischen Bewertung und Gestaltung von Bildschirmarbeit. Zürich, ETH.

N. Urbach, S. Smolnik, et al. (2009). "Der Stand der Forschung zur Erfolgsmessung von Informationssystemen. Eine Analyse vorhandener mehrdimensionaler Ansätze." Wirtschaftsinformatik **51**(4).

H. Willumeit, G. Gediga, et al. (1995). Validation of the IsoMetrics usability inventory, Universität Osnabrück.

G. Zülch and S. Stowasser (2002). "Bewertung der Gebrauchstauglichkeit von Software." Zeitschrift für Arbeitswissenschaft **56**(4).