Softwareevaluation in Gruppen oder Einzelevaluation: Sehen zwei Augen mehr als vier?

K.-C. Hamborg, G. Gediga, M. Döhl, P. Janssen & F. Ollermann

Fachbereich Psychologie, Universität Osnabrück

Zusammenfassung

Es finden sich Hinweise dafür, daß die Evaluation von Softwaresystemen effektiver in Gruppen- als in Einzelsettings ist. In drei Untersuchungen mit verschiedenen Varianten des IsoMetrics-Verfahrens werden diese Befunde überprüft. Die Ergebnisse zeigen Vorteile des Gruppensettings in Bezug auf die Qualität konstruktiver Anmerkungen, nicht jedoch in Bezug auf die Menge und inhaltliche Breite der Informationen und den Durchführungsaufwand. Spezifische Unterschiede der Methoden und Konsequenzen für die Praxis werden diskutiert.

1 Problemstellung

In diesem Beitrag wird ein möglicher Ansatz zur Effizienzsteigerung formativer Evaluation geprüft. Auf der Grundlage des IsoMetrics Fragebogens [10] wird die Einzelversion des Verfahrens mit einer gruppenbezogenen Variante verglichen. Ausgangspunkt für diese Untersuchung sind Befunde aus Vergleichsuntersuchungen von gruppen- und personenbezogenen Evaluationsmethoden, die darauf verweisen, daß sich die Effizienz von Evaluationsmethoden durch gruppenbezogene Settings steigern läßt. Zwei Arbeiten zu diesem Thema sollen im folgenden kurz dargestellt und durch allgemeine Argumente aus der Forschung von Gruppendiskussionsverfahren ergänzt werden.

Hackman und Biers [5] untersuchten die Methode des Lauten Denkens als Einzel- und als Gruppenmethode. Die Ergebnisse der Untersuchung zeigen, daß Evaluatoren im Zweier-Team insgesamt mehr Zeit mit dem Verbalisieren verbrachten als einzelne Evaluatoren. Im Team wurde mehr Zeit darauf verwendet, qualitativ hochwertige Verbalisierungen vorzunehmen. Wurden die Werte jedoch an der Personenzahl relativiert, gingen die Effekte verloren. Die Autoren halten die Ergebnisse, daß zwei Personen gemeinsam mehr Zeit mit dem Verbalisieren verbringen und dabei auch mehr hochwertige Informationen als eine Person generieren, für nicht trivial. Sie bewerten die Ergebnisse dahingehend, daß die Teammethode im gleichen Zeitraum mehr Informationen bringt als die Einzelmethode und daher zeiteffizienter funktioniert. Weiterhin verweisen sie darauf, daß im Team andere Anmerkungsqualitäten als bei der Einzelevaluation entstehen. Dazu zählen sie insbesondere auch Anmerkungen, die Unsicherheit mit dem Interface und dessen Bedienung ausdrücken.

In einer Untersuchung verglich Desurvire [1] Labortests mit Inspektionsmethoden. Ein System wurde mit der Methode des „Cognitive Walkthrough" und der „Heuristischen Evaluation" jeweils als Gruppen- und als Einzelmethode bewertet. Desurvire kommt zu dem Schluß, daß sowohl Experten als auch Nicht-Experten eine größere Anzahl Probleme, gemessen an dem Kriterium der Ergebnisse aus Labortests, vorhersagen, wenn die Evaluation in dem Gruppenparadigma durchgeführt wurde.

1.1 Vorteile gruppenorientierter Verfahren

Die Befunde der dargestellten Untersuchungen erscheinen aus methodischer Sicht plausibel. Teamorientierte Settings ähneln prinzipiell Gruppendiskussionsmethoden [2, 7]. Im Vergleich zu Einzelinterviews wird gruppenorientierten Verfahren zugeschrieben, daß sie einen größeren Bereich verschiedener Reaktionsweisen erfassen, Befragte zu detaillierten Meinungsäußerungen anregen und zur Aktualisierung und Explikation „tieferliegender Bewußtseinsinhalte" stimulieren, daß sie weiterhin zum Abbau psychischer Kontrollen beitragen und eher spontane, unkontrollierte Reaktionen, die den Schluß auf den latenten Inhalt geäußerter Meinungen zulassen, provozieren. Unter dem Eindruck der Auskunftsbereitschaft anderer Gesprächspartner sollten sie vor allem auch stärker gehemmte Teilnehmer zu Beiträgen ermutigen [2]. Weiterhin wird erwähnt, daß gruppenorientierte Verfahren auch ökonomischer seien. Ein geringerer sachlicher und personeller Aufwand erbringe soviel Material wie mehrere Einzeluntersuchungen [2, 7].

Die genannten Argumente beziehen sich in erster Linie auf den Vergleich von Einzelinterviews und Gruppendiskussionen. Nach unserer Auffassung gelten sie jedoch auch für den Vergleich schriftlicher Einzelbefragung und gruppenorientierter Befragungstechniken.

1.2 Mögliche Nachteile gruppenorientierter Verfahren

Jedoch wird auch eine Reihe möglicher Nachteile gruppenorientierter Verfahren im Vergleich zu Einzelinterviews genannt [2, 7]. Sie beziehen sich 1.) auf Beeinträchtigungen bei der Informationsgewinnung, 2.) auf gruppendynamische Prozesse und 3.) auf ökonomische Aspekte. Die Argumente werden im folgenden ebenfalls kurz dargestellt.

zu 1.) Relevante Informationen werden aus Gründen, die in der Person liegen, nicht genannt. Ursachen dafür können mangelnde Motivation oder Unkenntnis sein:

- Es liegt im Ermessen der TeilnehmerInnen, ob, wann und in welchem Umfang sie sich zu dem thematisierten Inhalt äußern. Es kann nicht ausgeschlossen werden, daß Personen zu einzelnen Punkten der Diskussion gänzlich schweigen oder Inhalten ausweichen.
- Einzelne Personen und Meinungen kommen nicht zum Zug, es kann zu einer ungleichmäßigen Beteiligung einzelner Personen und einer mengenmäßig unterschiedlichen Verteilung der Beiträge kommen.

zu 2.) Das Zusammenspiel einzelner Personen in der Gruppe wirkt sich auf den Informationsgewinn aus:

- Gruppendynamische Effekte können sich möglicherweise ungünstig auf Atmosphäre und Gesprächsbereitschaft auswirken.
- Gesprächsmonopolsierung, Gruppendruck, Beeinflussungsversuche oder die frühe Verständigung über Gruppenstandards bzw. der Rekurs auf bestimmte Normen können verhindern, daß ein breites Spektrum von Perspektiven zu einem Thema entwickelt wird.

zu 3.) Die Erstellung und Verarbeitung von Diskussionsprotokollen ist zeitaufwendig und mit vielfachen Schwierigkeiten struktureller und inhaltlicher Art verbunden. Die Auswertungsökonomie kann daher als eher ungünstig bewertet werden. Weiterhin wird auf den hohen organisatorischen Aufwand bei der Vorbereitung von Gruppendiskussion hingewiesen.

Die referierten Forschungergebnisse zeigen widersprüchliche Befunde zu Vor- und Nachteilen von gruppen- und einzelpersonbezogenen Befragungsverfahren, während aus dem direkten Anwendungsgebiet der Softwareevaluation vorteilhafte Effekte, insbesondere mit Bezug auf die Menge und die Qualität gewonnener Informationen, berichtet werden.

In dem vorliegenden Bericht wird das Verhalten von gruppen- und einzelpersonbezogenen Evaluationsmethoden am Beispiel entsprechender Varianten des IsoMetrics-Verfahrens untersucht. Der Vergleich bezieht sich nach *formativen* Gesichtpunkten auf die Menge, die Qualität und das Spektrum erhobener Bemerkungen zu einem Softwaresystem als Ausgangspunkt für Verbesserungen und nach *summativen* Gesichtspunkten auf die Profilbewertung des Programms aufgrund von Bewertungsskalen.

2 Methoden

In diesem Abschnitt werden die beiden Varianten des IsoMetrics- Verfahrens dargestellt.

2.1 IsoMetrics^{L}

Der Fragebogen IsoMetrics^{L} umfaßt sieben Subskalen mit derzeit insgesamt 75 Items. Jede Subskala von IsoMetrics^{L} repräsentiert einen Gestaltungsgrundsatz gemäß ISO 9241/10. Die Skalen sind den Gestaltungsgrundsätzen entsprechend benannt: Skala 'A': *Aufgabenangemessenheit*, Skala 'S': *Selbstbeschreibungsfähigkeit*, Skala 'T': *Steuerbarkeit*, Skala 'E': *Erwartungskonformität*, Skala 'F': *Fehlerrobustheit*, Skala 'I': *Individualisierbarkeit*, Skala 'L': *Erlernbarkeit*.

Je Item werden drei unterschiedliche Daten erhoben (s. Abbildung 1):

1. Die Bewertung der Software auf einer 5-stufigen Ratingskala in Bezug auf das Item. Die Skalierung der Skala reicht von "stimmt nicht" (Rating 1) bis "stimmt sehr" (Rating 5).
2. Die Gewichtung des Items als Ausdruck der Bedeutung des durch das Item operationalisierten Aspekts der ISO 9241/10 in Bezug auf den Gesamteindruck der Software. Die Gewichtung wird ebenfalls auf einer 5-stufigen Ratingskala, von "nicht wichtig" (Rating 1) über "mittelmäßig wichtig " (Rating 3) bis "sehr wichtig" (Rating 5) vorgenommen.
3. Eine oder mehrere auf das Item bezogene Problemmeldungen, die der oder die Befragte in freiem Text selbst formuliert.

Bearbeitet wird der IsoMetrics von prospektiven Nutzern des zu evaluierenden Systems. Die Konstruktion, Validität und Reliabilität des Verfahrens als auch die Gebrauchstauglichkeit der von den Benutzern erzeugten Bemerkungen für die Software-Entwicklung wurden in [10, 11] dokumentiert.

	stimmt nicht	Stimmt Wenig	Stimmt Mittelmäßig	stimmt ziemlich	stimmt sehr	Keine Angabe
Wenn Menü-Optionen in bestimmten Bearbeitungsschritten nicht zur Verfügung stehen, wird mir die Sperrung sichtbar gemacht.	1	2	3	4	5	

	nicht wichtig	Wenig Wichtig	Mittelmäßig Wichtig	ziemlich wichtig	sehr wichtig	Keine Angabe
Wie wichtig ist dieser Aspekt für Ihren Gesamteindruck von der Software?	1	2	3	4	5	

Können Sie konkrete Beispiele nennen, bei denen Sie dieser Aussage nicht zustimmen können?

Abbildung 1: Ein Fragebogenitem aus IsoMetricsL (Erläuterungen im Text)

2.2 IsoMetrics[G]

Als Modifikation bietet das IsoMetrics-Verfahren eine Gruppenversion an. Auch hier werden alle Items des Fragebogens benutzt, wobei jede Frage vom Befragungsleiter auf einem Overhead-Projektor vorgelegt wird. Jeder Teilnehmer erhält folgendes Format für die Beantwortung der Fragen (Abbildung 2).

		Stimmt nicht	Stimmt Wenig	Stimmt Mittelmäßig	stimmt ziemlich	Stimmt Sehr	Keine Angabe
A.1	Die Software zwingt mich Arbeitsschritte durchzuführen, die für meine Arbeit nicht sinnvoll sind.	1	2	3	4	5	

	Nicht wichtig	Wenig wichtig	Mittelmäßig wichtig	ziemlich wichtig	Sehr wichtig	Keine Angabe
Wie wichtig ist dieser Aspekt für Ihren Gesamteindruck von der Software?	1	2	3	4	5	

Kreuzen Sie bitte von denen auf der Overhead-Folie aufgelisteten Problemen diejenigen an, die Sie als relevant empfinden.

(1) (2) (3) (4) (5) (6) (7) (8) (9)

Abbildung 2: Ein Item aus dem IsoMetricsG-Fragebogen

Die Vorgehensweise in der Gruppenbefragung gestaltet sich wie folgt:

- Die Befragung der Gruppe wird von einem Moderator durchgeführt.
- Der Moderator benutzt pro IsoMetrics-Frage eine Overhead-Folie, trägt den Text des Kriteriums vor und erläutert u.U. diesen Text weiter – gibt Beispiele, etc.
- Die beteiligten Personen notieren jeder für sich das Rating für das Zutreffen der Aussage.
- Danach wird die Gewichtung von den Personen eingeschätzt (zweites Rating im Fragebogen).
- In der Gruppe werden Schwachpunkte der Software generiert, die der Moderator auf der Folie festhält.

- Die Schwachpunkte werden durchnumeriert.
- Am Ende der Generierung der Schwachpunkte für ein Item kreuzt jede beteiligte Person für sich an, welche der festgehaltenen Schwachpunkte ihrer Meinung nach relevant sind.

3 Versuchsplan und Methodik

3.1 Die Untersuchungen

Es wurden drei Untersuchungen durchgeführt, deren Gegenstand das Online Recherche Programm (OPAC) der Universitätsbibliothek Osnabrück war. In der ersten Untersuchung wurde das Programm mit IsoMetrics^{L} (N = 15), in der zweiten und dritten Untersuchung mit IsoMetrics^{G} (jeweils N = 7) bewertet. Die „Evaluatoren" waren Studierende der Universität Osnabrück, entstammten also der typischen Nutzerpopulation des Systems. Die Gruppen waren in Bezug auf Alter und Vorerfahrung nach der Anzahl genutzer Programme homogen. In Bezug auf die Vorerfahrung nach zeitlichen Aspekten unterscheiden sich die IsoMetrics^{L}- und die zweite IsoMetrics^{G} Gruppe, nicht aber die anderen Gruppen voneinander (siehe Tabelle 1). Die Evaluatoren verfügten über keine Vorbildung in Bezug auf Fragen der Software-Ergonomie.

Vor der Bewertung explorierten die Evaluatoren das Programm ca. 30 Minuten. Daraufhin bearbeiteten sie mit dem Programm vier Aufgaben. Für die Aufgabenbearbeitung standen weitere 15 Minuten zur Verfügung. Bei den Aufgaben handelte es sich um vier Rechercheaufträge, bezogen auf die Bestände der Universitätsbibliothek Osnabrück (s. Kasten 1).

1) Wie viele Einträge finden sich unter dem Titelstichwort „Organisationspsychologie"?
_____Treffer.

2 a) Schauen Sie sich den Eintrag zu Gros, Eckhardt: *Anwendungsbezogene Arbeits-, Betriebs- und Organisationspsychologie an*, der unter dem Titelstichwort „Organisationspsychologie" auftaucht.
b) Erkunden Sie, ob das Buch ausgeliehen ist.
c) Drucken Sie den Anzeigetext in Langdarstellung aus.
d) Übernehmen Sie den Titel in ein Speicherset.
e) Löschen Sie bitte das Speicherset.

3a) Suchen Sie alle Bücher mit dem Titelstichwort „Organisationspsychologie" von Siegfried Greif.
b) Speichern Sie die ersten drei Titel in einem Speicherset.

4a) Fügen Sie dem Speicherset die Arbeiten von Eberhard Ulich unter dem Titelstichwort „Arbeitspsychologie", die seit 1990 erschienen sind, hinzu.
b) Lassen Sie sich das gesamte Speicherset anzeigen.
c) Drucken Sie das Speicherset in Kurzdarstellung aus.

Kasten 1: Testaufgaben

Die Aufgaben wurden von den UntersuchungsteilnehmerInnen eigenständig bearbeitet. Die VersuchsleiterInnen konnten notfalls um Hilfe gebeten werden. Dies geschah jedoch nur in wenigen Fällen und überwiegend bei Problemen mit dem Telnet Programm, das Voraussetzung für die Arbeit mit dem OPAC-System war. Sowohl für die Explorationsphase als

auch für die Aufgabenbearbeitung konnten die Evaluatoren ein Kurzmanual, wie es Nutzern des Programms in der Regel zur Verfügung steht, benutzen.

	Alter	Vorerfahrung, Zeit (Index*)	Vorerfahrung, Anzahl genutzter Programme (*N*)			
			T	D	K	P
IsoMetrics^{L} (*N* = 15)	*M* = 25,4; *S* = 5,12	*M* =4,5; *S* = 1,13	12	3	4	1
IsoMetrics^{G} 1 (*N* = 15)	*M* = 23,6; *S* = 3,87	*M* =5,0; *S* = 1,0	7	1	1	1
IsoMetrics^{G} 2 (*N* = 15)	*M* = 24,6; *S* = 3,41	*M* =6,14; *S* = 1,4*	7	2	2	2

Erläuterungen:
M = Mittelwert; *SD* = Standardabweichung, T = Textverarbeitung, D = Datenbank, K = Kalkulation, P = Programmiersprache; Index = aggregiert aus Vorerfahrung in Jahren, Nutzungshäufigkeit in Tagen/Monat und Stunden/Sitzung; * Differenz (Scheffé) zu IsoMetrics^{L} p = ,021.

Tabelle 1: Charakterisierung der Untersuchungsgruppen nach Vorerfahrung und Alter

3.2 Abhängige Variablen und Datenaufbereitung

Mögliche Effekte der Methodenvarianten können nach formativen Gesichtspunkten bezüglich der Menge, der Qualität und des Spektrums der erhobenen Anmerkungen, sowie nach summativen Gesichtspunkten in Bezug auf die Bewertungsergebnisse der IsoMetrics-Skalen erwartet werden. Für die summative Bewertung wurden die Mittelwertsprofile und die Varianzen der Subskalen des IsoMetrics berechnet.

Voraussetzung für die Auswertung des Datenmaterials der formativen Analyse war als erster Schritt die Explikation [8] der durch die Evaluatoren vorgenommenen Anmerkungen zu dem Programm. Die Explikation konzentrierte sich darauf, die Anmerkungen gegebenenfalls so zu ergänzen, daß sie durch Dritte ohne zusätzliche Informationen verständlich waren, d.h. es wurden unvollständige Sätze ergänzt, Bezüge zu den Fragebogenitems hergestellt etc..

In einem nächsten Schritt wurden redundante, d.h. inhaltlich übereinstimmende Aussagen aus dem Datensatz herausgefiltert und gekennzeichnet. Für die unterschiedlichen Auswertungsschritte bildete der so aufbereitete Datensatz die Grundlage. Die weitere Datenauswertung wird aus Gründen der Übersichtlichkeit im Zusammenhang mit den Ergebnissen dargestellt.

4 Ergebnisse

4.1.1 Summative Bewertung

Grundlage für die summative Bewertung der Software durch IsoMetrics sind die Ausprägungen auf den Subskalen des Verfahrens. Abbildung 3 zeigt die Varianzen um die Mittelwerte der Ratings für die drei Untersuchungen. In Abbildung 4 sind die aus den Subskalenwerten resultierenden Mittelwertsprofile für IsoMetrics^{G} aus den Untersuchungen dargestellt und in Abbildung 5 ist das Mittelwertsprofil der zusammengefaßten Werte von IsoMetrics^{G} zu dem Profil aus der IsoMetrics^{L} Untersuchung in Beziehung gesetzt.

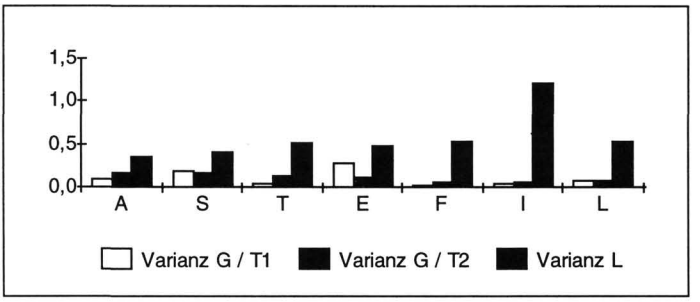

Abbildung 3: Varianzen der Ratings von IsoMetricsG und IsoMetricsL

Es ist zu bemerken, daß die Bewertungen der Einzelversion deutlich stärker streuen als die der Gruppenversion. Der Vergleich der Werte der beiden Untersuchungen mit IsoMetrics^{G} (Abbildung 4) zeigt bei relativ geringen Skalenunterschieden (mit Ausnahme der Skala T) bedeutsame Unterschiede (Effektgröße > 0,8) außer bei der Skala Individualisierbarkeit (I). Aus statistischer Sicht wird dies durch die geringe Streuung der Skalenwerte der IsoMetrics^{G} Untersuchungen begünstigt (s. Abbildung 3).

Der Vergleich der Bewertungsergebnisse aus der Gruppen- mit denen der Einzelversion (Abbildung 5), zeigt, daß die Gruppenversion zu einer „strengeren" Bewertung als die Einzelversion führt. Der Unterschied der Bewertungsergebnisse ist auf allen Subskalen signifikant und bedeutsam (Effektgröße > 0,8). Es ist jedoch hervorzuheben, daß sich die Profilverläufe nicht strukturell unterscheiden: Werden die aggregierten Mittelwertsprofile aus den Untersuchungen mit der Gruppenversion um einen Skalenpunkt erhöht, unterscheiden sie diese nicht mehr von dem Profil, das aus der Einzeluntersuchung resultiert.

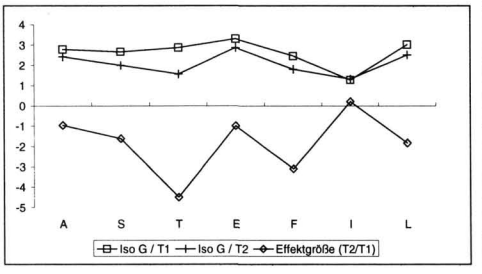

Abbildung 4: Mittelwertprofile aus IsoMetrics^{G}, Untersuchung 1 und 2

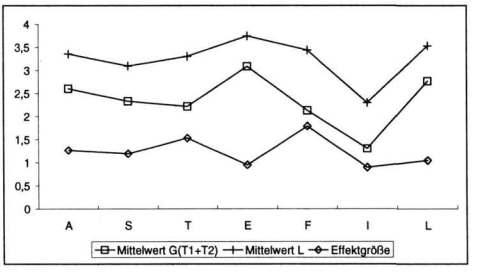

Abbildung 5: Mittelwertprofile IsoMetrics^{G} und IsoMetrics^{L}

Zusammenfassung

Die Ergebnisse der Gruppenversion zeigen im Vergleich zu der Einzelversion eine deutlich geringere Varianz bei der summativen Bewertung des zu evaluierenden Systems. Dies kann als Indikator dafür gewertet werden, daß durch den Gruppenprozeß extreme Bewertungen ausgeschlossen werden und eine Art Gruppenstandard homogenisierende Wirkung zeigt.

Weiterhin ist zu beobachten, daß die Gruppenversion zu einer strengeren Beurteilung des Systems führt. Die strengere Beurteilung erfolgt systematisch, d.h. die zusammengefaßten Profile aus der Gruppenversion und der Einzelversion unterscheiden sich um eine Konstante, unabhängig von der Wahl des Kriteriums.

4.1.2 Formative Bewertung

4.1.2.1 Vergleich der Anzahl erhobener Anmerkungen

Von den 15 Evaluatoren, die mit IsoMetrics^{L} arbeiteten, wurden insgesamt 345 Anmerkungen zu dem System formuliert.

Für die weitere Auswertung wurden davon 35 Bemerkungen nicht berücksichtigt. Dies waren Bemerkungen, die sich auf den Fragebogen (z.B. dessen Verständlichkeit) bezogen, (3), Bemerkungen, die sich auf das Betriebssystem oder auf die Programmumgebung richteten (11), oder die auf andere Bemerkungen verwiesen (9) sowie unvollständige, nicht explizierbare oder unverständliche Anmerkungen (12). Die resultierenden 310 Items enthielten 65 redundante, d.h. thematisch übereinstimmende Anmerkungen, die in der weiteren Auswertung ebenfalls nicht berücksichtigt wurden. Es verblieben also insgesamt 245 redundanzfreie und weiterhin auswertbare Anmerkungen. Das entspricht im Durchschnitt 16,3 Anmerkungen pro Teilnehmer der Evaluation.

In der ersten Untersuchung mit der Gruppenversion des IsoMetrics wurden 129, in der zweiten 110 Anmerkungen erhoben. Darin sind in der ersten Untersuchung 37, in der zweiten 29 Redundanzen enthalten. Ohne diese Redundanzen verblieben für die erste Untersuchung 92 und für die zweite 81 Anmerkungen. Durchschnittlich sind dies 13,1, respektive 11,6 Anmerkungen pro Evaluator.

Der Häufigkeitsvergleich der mit der Gruppen- und der Einzelversion erhobenen Anmerkungen zeigt, daß durch den Einsatz der Einzelversion mehr Anmerkungen (Chi^2 = 12,4, df=1, p<1%) erhoben wurden als durch die einzeln eingesetzte Gruppenversion, aber auch im Verhältnis zu den aus G1 und G2 (Kürzel ISO-G) aggregierten Ergebnissen (s. Tabelle 2).

Methode	Anmerkungen ursprünglich	Ausgesonderte Anmerkungen	Redundanzen	Auswertbare Anmerkungen ohne Redundanzen
L	345	35	65	245
ISO-G	239		66	173
G1	129		37	92
G2	110		29	81

Tabelle 2: Anmerkungshäufigkeiten IsoMetrics L und G

Zusammenfassung

Die Annahme, daß durch gruppenbezogene Evaluationsverfahren mehr Anmerkungen erhoben werden, bestätigt sich in der vorliegenden Untersuchung bei etwa gleicher Anzahl Evaluatoren nicht.

4.1.2.2 Vergleich der Anmerkungen nach Relevanz

Zur Bestimmung der Relevanz der erhobenen Anmerkungen wurden die von den Untersuchungsteilnehmern erzeugten redundanzfreien Anmerkungen zunächst verschiedenen für Entwickler relevanten Kategorien zugewiesen (Tabelle 3, s.a. [3]).

Kategorie	Art der Bemerkung
Irr	Die Aussage bezieht sich nicht auf die Software.
Pos	Die Aussage ist positiv und zeigt keine Schwächen des Systems auf.
X	Die Bemerkung drückt ein generelles Unverständnis aus. Sie handelt mehr vom Fühlen, Denken und Erleben der Person als von den auslösenden Schwächen der Software. Es muß aus der Bemerkung ersichtlich sein, daß es sich bei der Software um den auslösenden Faktor handelt.
S	Die Aussage bezieht sich auf eine lokalisierbare Stelle oder eine bestimmte Funktion oder einen bestimmten Befehl innerhalb der Software. Der Nutzungskontext muß in der Anmerkung genannt sein.
S+	Wie Kategorie S, die Aussage enthält jedoch mindestens einen konkreten Vorschlag ***wie*** das Problem behoben werden könnte.
G	Die Aussage bezieht sich auf ein Problem, das die gesamte Software durchzieht. Es ist nicht genau angegeben, welche Stellen betroffen sind. Beispiele, die zur Illustration dienen, machen aus einer generellen Anmerkung keine spezielle Anmerkung.
G+	Wie Kategorie G, die Aussage enthält jedoch zusätzlich mindestens einen konkreten Vorschlag ***wie*** das Problem behoben werden könnte.

Tabelle 3 Kategoriendefinition für die Analyse der Bemerkungen

Die Zuordnung der Anmerkungen zu den Kategorien erfolgte durch 4 Rater. Drei der Rater waren Studierende der Universität Osnabrück, Fachbereich Psychologie, die spezielle Kurse zu den Themen Evaluation und Software-Ergonomie besucht hatten, der vierte Rater ist Lehrender in diesem Gebiet. Vor dem Rating wurde ein Trainingsdurchgang durchgeführt. Bei dem Training und dem nachfolgenden Rating standen ein Informationsblatt mit Erläuterungen zu den Kategorien und mit Abgrenzungsbeispielen zur Verfügung.

Für die Auswertung der in diesem Abschnitt verfolgten Frage wurden nur die Anmerkungen berücksichtigt, die von wenigstens drei Ratern übereinstimmend kategorisiert wurden. Nach diesem Kriterium konnten 356 der insgesamt 418 redundanzfreien Anmerkungen (ca. 85%) ausgewertet werden. Neben der Tatsache, daß bei Kategorisierungen so gut wie nie eine völlige Übereinstimmung erreicht wird, können einige allgemein formulierte Anmerkungen, die eine trennscharfe und übereinstimmende Zuordnung zu den Kategorien teilweise erschwerte, als Ursache für die fehlende Übereinstimmung der Rater in 15% der Fälle gesehen werden.

Tabelle 4 zeigt die Häufigkeitsverteilung dieser Anmerkungen zu den Kategorien in Abhängigkeit von der eingesetzten Methodenvariante.

Kategorie	G1		G2		ISO-G		L	
	N	%	N	%	N	%	N	%
Irr, Pos, X	0 (-)	0 %	1 (-)	1,4%	1 (-)	,7%	27 (+)	12,7%
S	26 (+)	33,8%	14	20,0%	40 (+)	28%	29 (-)	13,6%
S+	3	3,9%	1	1,4%	4	2,8%	5	2,3%
G	47	61%	51	72,9%	94	65,7%	145	68,1%
G+	1	1,3%	3	4,3%	4	2,8%	7	3,3%
Gesamt	77	100 %	70	100 %	143	100 %	213	100 %

Tabelle 4: Häufigkeitsverteilung der Anmerkungen zu den Relevanzkategorien

Nach Berechnung einer Konfigurationsfrequenzanalyse zeigen die mit (+) gekennzeichneten Werte bedeutsam positive bzw. die mit (-) gekennzeichneten Werte auf dem 5% Niveau bedeutsam negative Abweichungen von dem erwarteten Wert aus dem Vergleich von IsoMetrics^{L} und IsoMetrics^{G} (Untersuchung G1, G2, bzw. ISO-G gegen Untersuchung L). Die Verteilung der Anmerkungen auf die unterschiedlichen Relevanzklassen zeigt, daß der Anteil irrelevanter, positiver und der Anmerkungen aus Kategorie X bei der Einzelversion bedeutsam überwiegt, während die Gruppenversion verhältnismäßig mehr spezielle Anmerkungen in Untersuchung G1 als auch bei Berücksichtigung der Anmerkungen aus ISO-G (Kategorie S), jedoch nicht in Untersuchung G2, evoziert.

Zusammenfassung

Durch die Gruppenvariante des Verfahrens werden weniger nicht konstruktiv verwendbare Anmerkungen erhoben. Weiterhin zeigen sich Unterschiede bei speziellen Anmerkungen (Kategorie S). Durch die Gruppenversion wurden in Untersuchung G1 und bei Berücksichtigung der aggregierten Anmerkungen aus G1 und G2 mehr Anmerkungen dieser Kategorie als mit der Einzelversion erhoben.

4.1.2.3 Vergleich des Anmerkungsspektrums

Weiterhin wurde in der vorliegenden Untersuchung das Anmerkungsspektrum in Abhängigkeit von der Methodenvariante untersucht. Ausgangspunkt für diese Analyse war die Verteilung der Anmerkungen auf die Subskalen des IsoMetrics.

Hierzu wurden die redundanzfreien Anmerkungen pro Skala für die Methodenvarianten ausgezählt. Die Ergebnisse sind in Tabelle 5 dargestellt. Sie zeigen, daß sich die mit IsoMetrics^{L} und IsoMetrics^{G} erhobenen Anmerkungen relativ zu der Gesamtmenge unterschiedlich auf die Subskalen verteilen (Chi^2=38,7, df=6, p<1%). Die mit (+)/(+*) und (-)/(-*) gekennzeichneten Werte markieren bedeutsam positive bzw. negative Abweichungen vom erwarteten Wert auf dem 1% bzw. 5% Niveau.

	Aufgabenangemessenheit	Selbstbeschreibungsfähigkeit	Steuerbarkeit	Erwartungskonformität	Fehlertoleranz	Individualisierbarkeit	Erlernbarkeit	gesamt
ISO-G	103 (+)	23	15	8 (-*)	16	0 (-)	8 (-)	173
L	79 (-)	43	29	22 (+*)	30	16 (+)	26 (+)	245
Gesamt	182	66	44	30	46	16	34	418

Tabelle 5: Häufigkeitsverteilung der Anmerkungen über die Subskalen des IsoMetrics

In der Gruppenversion finden wir eine starke Konzentration auf die erste Skala (Aufgabenangemessenheit) bei weitgehendem Abfall auf den folgenden Subskalen. Bei der Einzelversion werden ebenfalls die meisten Anmerkungen durch die erste Skala (Aufgabenangemessenheit) erhoben, die folgenden Skalen tragen aber in einem stärkeren Maße als bei IsoMetrics[G] zur Erhebung weiterer Anmerkungen bei.

Zusammenfassung

Die Anmerkungen der gruppenorientierten Variante konzentrieren sich auf die zuerst bearbeitete Skala (Aufgabenangemessenheit). Vermutlich wurde die Auskunftbereitschaft der Evaluatoren für die folgenden Skalen dadurch weitgehend erschöpft. Hierdurch ensteht die Gefahr, daß sich die Evaluationsergebnisse auf ein eingeschränktes Spektrum der zugrundegelegten Gestaltungsgrundsätze reduziert. Bei der Einzelversion des Fragebogens verteilen sich die Anmerkungen dagegen auch auf die in der Bearbeitungsfolge später dargebotenen Skalen.

4.1.2.4 Durchführungsaufwand

Die Einzel- und die Gruppenversion unterscheiden sich nach ihrem Durchführungs- und Auswertungsaufwand.

In Bezug auf die Durchführung erfordert die Gruppenversion mehr organisatorische Vorbereitungen, da die TeilnehmerInnen auf einen Termin hin koordiniert werden müssen. Hier ermöglicht die Einzelversion mehr Flexibilität, da prinzipiell Personen unabhängig voneinander den Fragebogen bearbeiten können. Da für die Durchführung der Gruppenversion ein geschulter Moderator und zusätzlich ein Protokollant, für die Aufzeichnung der Anmerkungen der TeilnehmerInnen, benötigt werden, ist nach personellen Gesichtspunkten diese Verfahrensvariante aufwendiger. Die Untersuchung mit der Einzelversion kann dagegen von einer angelernten Hilfskraft durchgeführt werden. Der Zeitaufwand für die Gruppenversion betrug in den hier dargestellten Untersuchungen ca. 3 Stunden, für die Durchführung der Einzelversion 1,5 bis max. 2 Stunden (inkl.Pausen, ohne Explorationsphase und Aufgabenbearbeitungen).

Der Auswertungsaufwand der summativen Ergebnisse ist versionsspezifisch nicht unterschiedlich. Dies ist jedoch anders bei der Auswertung der formativen Ergebnisse. Bei den Ergebnissen der Gruppenversion besteht weniger Explikationsaufwand, da die Anmerkungen durch den Prozeß der Informationsvermittlung, der die Verbalisierung und Niederschrift der Anmerkungen in einer für alle GruppenteilnehmerInnen verständlichen Weise umfaßt, zum größten Teil besser ausformuliert sind. Ebenfalls ist die Auseinandersetzung mit nicht relevanten Aussagen bei der Gruppenvariante geringer als bei der Einzelvariante.

Zusammenfassend ist ein höherer Durchführungsaufwand nach Zeit- und Personalkosten für die Gruppenvariante festzuhalten, der insgesamt die Ersparnisse bei der Auswertung übersteigen dürfte.

5 Diskussion

In der vorliegenden Untersuchung wurde die Gruppen- und Einzelversion des IsoMetrics, eines Verfahrens zur formativen und summativen Evaluation von Software, untersucht. Ausgangspunkt für die Untersuchung waren Befunde, die auf den effizienzsteigernden Charakter gruppenorientierter Verfahren zur Evaluation von Software hinweisen.

Der Methodenvergleich zeigt, daß auf der summativen Ebene die Bewertungsergebnisse der Gruppenvariante (IsoMetrics^G) weniger streuen und um einen Skalenpunkt strenger bewerten als die der Einzelvariante (IsoMetrics^L). Der letztgenannte Aspekt ist insbesondere bei der vergleichenden Untersuchung von Softwaresystemen zu beachten, um den Einfluß methoden–spezifischer Bewertungstendenzen kontrollieren zu können.

Auf der formativen Bewertungsebene zeigte sich, daß mit der Einzelvariante (IsoMetrics^L) insgesamt mehr Anmerkungen zu der evaluierten Software erhoben wurden und die erhobenen Anmerkungen die Gestaltungsgrundsätze der zugrundeliegenden Norm breiter abdecken. Demzufolge können die Befunde von Desurvire [1], nach denen gruppenorientierte Evaluationsvarianten mehr Informationen liefern als einzelpersonorientierte, nicht gestützt werden.

Die Stärke der Gruppenvariante besteht darin, daß mehr spezielle Anmerkungen erhoben werden, diesbezüglich die Anmerkungsqualität im Unterschied zu der Einzelversion also besser ist. Dieser Effekt stimmt mit den in der Literatur genannten Befunden überein, daß Gruppendiskussionsverfahren zu detaillierteren Meinungsäußerungen anregen.

Ein weiterer Vorteil der Gruppenversion kann darin gesehen werden, daß weniger nichtverwertbare Informationen erhoben werden. Eine Erklärung hierfür ist, daß es in der Gruppensituation erforderlich ist, die Anmerkungen für die weiteren Mitglieder und den Moderator verständlich formulieren zu müssen. Die Hemmschwelle, wenig aussagekräftige Hinweise zu formulieren, dürfte in der Gruppensituation höher sein als bei der alleinigen Bearbeitung eines Fragebogens. Zum geringeren Anteil nicht verwertbarer Anmerkungen dürfte ebenfalls beitragen, daß diese in der Gruppenversion durch den Moderator weiterverarbeitet werden, wenn er die Aussagen auf Overheadfolie festhält, ggf. nachfragt und reformuliert.

Es kann zusammengefaßt werden, daß die untersuchten Methodenvarianten weder bei der summativen noch bei der formativen Evaluation von Software äquivalent funktionieren. Die summativen Bewertungsergebnisse der Gruppenvariante sind strenger und streuen weniger. Befunde in bezug auf einen besseren Mengenertrag von Information zur Optimierung evaluierter Programme durch gruppenbezogene Verfahren können nicht gestützt werden. In Bezug auf die Informationsqualität zeigen sich teilweise Vorteile der Gruppenvariante, nicht jedoch in Bezug auf das thematische Anmerkungsspektrum und den Durchführungsaufwand.

Vor dem Hintergrund der vorliegenden Ergebnisse stellt sich die Frage, welche Variante in der Praxis bevorzugt werden sollte. Der geringere Durchführungsaufwand, die größere Anmerkungsmenge und die inhaltliche Anmerkungsbreite sprechen für die Einzelversion, für die Gruppenversion die höhere Anmerkungsqualität, die wesentlich darin besteht, daß die Anmerkungen konkreter auf Funktionen der untersuchten Software gerichtet sind.

Verschiedene Autoren empfehlen, im Anschluß an die Systembewertung Usability Reviews, an denen ggf. Anwender, Nutzer und Systementwickler beteiligt sind, durchzuführen [6, 11, 12]. Hier lassen sich erhobene Funktions- und Gestaltungsprobleme an Hand vorsortierer Anmerkungen aus Evaluationsuntersuchungen priorisieren und Maßnahmen für das (Re-) Design planen [4, 11].

Im Rahmen solcher Reviews ließe sich auch die niedrigere Spezifität der Anmerkungen aus IsoMetrics^L kompensieren. Diesem Vorgehen folgend, wäre die Einzelvariante die effektivere und zu empfehlende Methode.

Der Einsatz des gruppenorientierten Settings ist jedoch dann angezeigt, wenn aufgrund von Vorbildung oder Erfahrung die Anwender des Verfahrens individuelle Unterstützung im Um-

gang mit den Bewertungsskalen und der Generierung von Problempunkten benötigen. Unter diesen Bedingungen ist die Gruppenvariante die effizientere und vorzuziehende Methode.

6 Literatur

[1] Desurvire, H. (1994). Faster, Cheaper!! Are Usability Inspection Methods as Effective as Empirical Testing?. In: J. Nielsen & R.L. Mack (eds.). *Usability Inspection Methods*. New York: J. Wiley & Sons.

[2] Dreher, M. & Dreher, E. (1994). Gruppendiskussion. In: G.L. Huber & H. Mandl (Hrsg.) *Verbale Daten. 2. Auflage*. Weinheim: Psychologie Verlags Union.

[3] Gediga, G. & Hamborg, K.-C. (1997). Heuristische Evaluation und IsoMetrics: Ein Vergleich. In: R. Liskowsky, B.M. Velichkovsky & W. Wünschmann (Hrsg.). *Software Ergonomie '97, Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung*. Stuttgart: Teubner.

[4] Gediga, G. & Hamborg, K.-C. (1997). *Das IsoMetrics-Manual*. Osnabrücker Schriftenreihe Software-Ergonomie 2.

[5] Hackmann & Biers (1992).Team usability testing: Are two heads better than one? *Proceedings of the 36th annual meeting of the Human Factors society*, 36, S. 1205-1209.

[6] Karat, C.-M. (1994). A Comparison of User Interface Evaluation Methods. In: J. Nielsen & R.L. Mack (eds.). *Usability Inspection Methods*. New York: J. Wiley & Sons.

[7] Mangold, W. (1973). Gruppendiskussionen. In. R. König (Hrsg.). *Handbuch der empirischen Sozialforschung, Bd. 2. Grundlegende Methoden und Techniken der empirischen Sozialforschung, Erster Teil*. Stuttgart: Enke.

[8] Mayring, P. (1997). *Qualitative Inhaltsanalyse. 6., durchgesehene Auflage*. Weinheim: Beltz Deutscher Studienverlag.

[9] Nielsen, J. (1993). *Usability Engineering*. Boston, AP Professional.

[10] Gediga, G., Hamborg K.-C. & Düntsch, I. (in print). The IsoMetrics Usability Inventory: An operationalisation of ISO 9241-10. *Behaviour and Information Technology*.

[11] Willumeit, H., Gediga, G. & Hamborg, K.-C (1996) IsoMetrics^{L}: Ein Verfahren zur formativen Evaluation von Software nach ISO 9241/10. *Ergonomie & Informatik*, 27, 5 - 12.

[12] Wixon, D., Jones, S., Tse, L. & Casady, G. (1994). Inspections and Design Reviews: Framework, History and Reflection. In: J. Nielsen & R.L. Mack (eds.). *Usability Inspection Methods*. New York: J. Wiley & Sons.

Adressen der Autoren

Dr. Günther Gediga,
Universität Osnabrück,
FB Psychologie
Fachgebiet Methodenlehre
Seminarstr. 20
49069 Osnabrück

Dr. Kai-Christoph Hamborg
Cand. Psych. Meike Döhl
Cand. Psych. Philip Janssen
Cand. Psych. Frank Ollermann
Universität Osnabrück,
FB Psychologie
Fachgebiet Arbeits- und Organisationspsychologie
Seminarstr. 20
49069 Osnabrück