Internet-Based Measurement With Visual Analogue Scales: An Experimental Investigation

Dissertation

der Fakultät für Informations- und Kognitionswissenschaften

der Eberhard-Karls-Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Frederik Funke

aus Mainz

Tübingen

2010

Erstgutachter: Prof. Dr. Ulf-Dietrich Reips

Zweitgutachterin: Prof. Dr. Ulrike Cress

Contents

List of Figures

List of Tables

Acknowledgments

The thesis on hand presents the main results of my methodological research as postgraduate. This preface, however, is not only about summarizing last years' professional efforts but also about a personal development. I therefore want to seize the opportunity to express my gratitude to all those who accompanied me on this way.

Academic recognition

First of all, many thanks to Prof. *Ulf-Dietrich Reips* who proved to be an excellent supervisor. He managed to bridge the geographical distance between his place of work – the University of Deusto, Spain, formerly the University of Zurich, Switzerland – and my location in Kassel, Germany. Our email communication exceeded 2,000 messages by far, an indication of close cooperation. Since 2005, we gave 16 joint paper presentations and 8 poster presentations at 19 international conferences in 11 countries (see Appendices A and B). Furthermore, we published one chapter in an edited book as well as two journal articles. Two additional manuscripts have been submitted for publication in 2010 that are currently under review. Thanks also for the professional experience I have been gaining since 2006 as an editorial assistant for the *International Journal of Internet Science* as well as during my short stay at the University of Zurich in 2007. I am also grateful to Prof. *Ulrike Cress* from the Institut für Wissensmedien (IWM, Knowledge Media Research Center) in Tübingen for being the second advisor of this dissertation on short notice.

While doing this PhD thesis, I was given a lot of support by members of the University of Kassel. Grateful thanks to Prof. *Marek Fuchs* (now at the Technical University of Darmstadt, Germany) and his assistants for many inspiring discussions about survey methodology and for the experience I gained during joint research that resulted in two publications. I would also like to thank Dr. *Manuela Pötschke,* deputy of the Department of Sociology, for lending a helping hand and for our discussions about statistical challenges. Furthermore, many thanks to Prof. *Ernst-Dieter Lantermann* (Department of Psychology) for giving me the opportunity to conduct Web experiments in his classes.

I would also like to thank Dr. *Randall K. Thomas* (ICF International, USA) for joint research and Prof. *Mick P. Couper* (University of Michigan, USA) for many discussions about VASs. The tool that was used to create the visual analogue scales for most of my research is the

free Web service VAS Generator, maintained by Ulf-Dietrich Reips and myself. I am very grateful, though, to Dr. *Timo Gnambs* (University of Linz, Austria) who developed the underlying software, based on an open source script by *Walter Zorn*.

In addition, I would like to extend my thanks to all scholars who gave helpful comments at conferences and in personal communication. Mentioning all by name would go beyond the scope of these acknowledgements. Thanks, therefore, to all participants I met in person during the fruitful *International Survey Methodology* (ISM) workshops in Lillehammer (2007) and Bergamo (2009), the conferences of the *European Survey Research Association* (ESRA) in Prague (2007) and Warsaw (2009), the conferences of the *Research Committee 33 "Logic and Methodology"* for the International Sociological Association (ISA-RC33) in Naples (2008) and Gothenburg (2010), the *XXIX International Conference of Psychology* in Berlin (2008), the *1st International Workshop on Survey Methodology* in Daejeon (South Korea, 2009), and to all discussants at the *General Online Research* (GOR) conferences in Zurich (2005), Bielefeld (2006), Leipzig (2007), Hamburg (2008), Vienna (2009), and Pforzheim (2010). Also thanks to *Deutscher Akademischer Austauschdienst* (DAAD, German Academic Exchange Service) for travel funding to the ESRA conference (2009) and the ISA-RC33 conference (2010). I would also like to thank for the academic experience I could gain as a member of the international board of the GOR conference and as reviewer for *Public Opinion Quarterly* and the *International Journal of Internet Science*.

Personal recognition

Last but not least, warmest thanks to my friends and family who supported me in various ways during the last years. On the part of the family, *Elke Funke-Kirste*, *Bernd Funke*, and *Agata Funke*. On the part of the friends, special thanks to *Christine Nowack*, *Sebastian Koch, Christian Reuter*, and *Benjamin Pilz*, as well as *Okkinara*, *Katami*, and *Lina* for their awesome support.

Kassel, August 2010
Frederik Funke

i

Zusammenfassung

(German summary)

i Zusammenfassung (German Summary)

i.i Einführung und Rahmen

i.i.i Psychologische Forschung im Internet

Das Internet als Medium akademischer Forschung ist längst keine Randerscheinung mehr (siehe Batinic, Reips, & Bosnjak, 2002; Gosling & Johnson, 2010; Reips 2006b, 2006c; Reips & Bosnjak, 2001; Sassenberg, Boos, Postmes, & Reips, 2003). Auch in der Psychologie hat Forschung im World Wide Web einen festen Platz in Curricula und im Methodenkanon (Barak, 1999, 2008; Batinic, 2000; Döring, 2003; Joinson, McKenna, Postmes, & Reips, 2007; Reips & Birnbaum, in press) und ergänzt traditionelle Forschungsmethoden (Birnbaum, 2000a; Birnbaum & Reips, 2005; Eid & Diener, 2006; Reips, 2003). Skitka und Sargis (2006) halten in einer Metaanalyse fest, dass bereits zwischen 2003 und 2004 ein Fünftel der von der American Psychological Association (APA) gelisteten Zeitschriften mindestens einen Artikel veröffentlicht hatten, in dem Ergebnisse webbasierter Forschung berichtet wurden.

Wird das psychologische Labor vor Ort durch eine Zugangsmöglichkeit via Internet erweitert, spricht man von einem Weblabor (vgl. Birnbaum, 2000a; Musch & Reips, 2000; Reips, 2001; Skitka & Sargis, 2006). Bei der Forschung im Weblabor entfällt die Notwendigkeit der physischen Gleichzeitigkeit von Labor, Versuchsleiter und Untersuchungsteilnehmer. An Webexperimenten kann grundsätzlich 24 Stunden am Tag an 365 Tagen im Jahr teilgenommen werden, solange ein Zugang zum Internet besteht. Dadurch werden Forschungsmöglichkeiten nicht nur quantitativ, sondern vor allem auch qualitativ erweitert (Janetzko, Hildebrand, & Meyer, 2003; McGraw, Tew, & William, 2000; Reips, 2000, 2001, 2002b, 2002c; Reips & Krantz, 2010). Computer kommen in der experimentellen Forschung bereits seit den 1970er Jahren zum Einsatz (Connes, 1972; Hoggatt, 1977), das Internet wird seit Mitte der 1990er Jahre für psychologische Forschung genutzt (siehe Reips, 2002b). Im Web durchgeführte Studien tragen zwangsläufig viele Charakteristika computerisierter Datenerhebung und computervermittelter Kommunikation (siehe Mühlenfeld, 2004; Reips, 2006a), was sowohl Vorteile als auch Nachteile und Risiken mit sich bringt, die in Kapitel 1 der vorliegenden Arbeit im Detail behandelt und im Folgenden nur kurz angesprochen werden sollen (siehe aber auch Best & Krueger, 2004; Couper, 2008; Krosnick, 1999; Reips, 2006b; Reips & Bosnjak, 2001; Schmidt, 1997; Welker & Wenzel, 2007).

Vorteile der Forschung im Web. Zu den *allgemeinen* Vorteilen computervermittelter Forschung zählen neben einer hohen Standardisierung die überwiegende Automatisierung des Untersuchungsprozesses, das sichere Durchführen anspruchsvoller Experimente, komplexe Filterführung, die konditionale Präsentation multimedialer Stimuli (Fuchs, 2008; Fuchs & Funke, 2007), sowie die Randomisierung von Frage- und Antwortreihenfolge. Hinzu kommt die Möglichkeit des Aufzeichnens von Prozessdaten wie Reaktionszeiten und Eingabekorrekturen, was Inferenz auf zugrunde liegende kognitive Prozesse erlaubt (siehe Heerwegh, 2003; Stieger & Reips, in press). In Untersuchungen im Web lassen sich Fehlerquellen wie Versuchsleitereffekte und soziale Erwünschtheit verringern (siehe Richman, Kiesler, Weisband, & Drasgow, 1999). Studien im Web weisen im Vergleich zur Laborsituation vor Ort zwar eine geringere technische Standardisierung auf, was einerseits die Gefahr des Auftretens technischer Probleme birgt. Andererseits kann so die externe Validität von Ergebnissen durch größere Varianz der Befragungssituation gesteigert werden. Letztlich entfällt die bei Papierfragebögen zeitlich und finanziell aufwändige und fehleranfällige Dateneingabe, sodass die Datenanalyse bereits parallel zur Feldphase erfolgen kann.

Die *speziellen* Vorteile von Untersuchungen im World Wide Web im Vergleich zu computerunterstützter Forschung vor Ort sind vor allem die schnelle Erreichbarkeit vieler, auch geographisch weit voneinander entfernter Teilnehmer. Zudem ermöglicht das Internet den Zugang zu speziellen und seltenen Populationen (z. B. Mangan & Reips, 2007; Marcell & Falls, 2001). Durch die immer größere Verbreitung von Internetzugängen und Daueranbindungen ans Netz, nicht nur zu Hause oder am Arbeitsplatz, sondern auch unterwegs durch den mobilen Zugang mit Mobiltelefonen und ähnlichen Geräten, werden weitere Möglichkeiten eröffnet (siehe Fuchs, 2008; Peytchev & Hill, 2010). Letztlich zeichnen sich Onlinebefragungen vor allem bei Untersuchungen mit einer großen Anzahl von Befragten durch geringere finanzielle Kosten aus.

Nachteile und Risiken der Forschung im Web. Grundsätzliche Nachteile computerisierter Untersuchungen sind vor allem mit der technischen Abhängigkeit von Hard- und Software verbunden. Auch wenn Onlinefragebögen nicht manuell programmiert werden müssen und zahlreiche Programme für deren Erstellung zur Verfügung stehen (z. B. Göritz & Birnbaum, 2005; Reips & Neuhaus, 2002), sollten Forscher über eine gewisse technische Expertise verfügen, um bestehende Möglichkeiten vollständig nutzen, die Struktur von Fragebögen überprüfen und Fallstricke erkennen zu können. Selbst bei größtmöglicher Sorgfalt und Berücksichtigung der Erkenntnisse der Umfragemethodologie (z. B. de Leeuw, Hox, & Dillman, 2008; Dillman, Smyth, & Christian, 2009; Groves et al., 2009) kann der Modus der Datenerhebung (z. B. Telefon versus Papier versus Web) einen großen Einfluss auf die je-

weils gewonnen Daten haben. Wenn verschiedene Erhebungsarten in einem Mixed-mode-Design (siehe Dillman & Smyth, 2007; Dillman et al., 2009; Fass & Schoen, 2006; Schonlau, Asch, & Du, 2003) parallel zum Einsatz kommen, kann der Erhebungsmodus als unabhängige Variable genutzt werden. Somit können Rückschlüsse darauf gezogen werden, welchen Einfluss das Medium Internet zum Beispiel auf psychologische Messungen haben kann (siehe auch Gosling & Johnson, 2010; Krantz & Dalal, 2000; Proctor & Vu, 2005; Reips & Bosnjak, 2001).

Ein weiteres ernst zu nehmendes Problem ist, dass sich der technischen Varianz auf Seiten der Untersuchungsteilnehmer nicht immer Rechnung tragen lässt. Da technische und psychologische Variablen konfundiert sein können (Buchanan & Reips, 2001), besteht das Risiko, dass Befragte mit bestimmten Charakteristika systematisch aufgrund technischer Restriktionen von Untersuchungen ausgeschlossen werden und Ergebnisse somit verzerrt werden (siehe auch Kapitel 5).

Wie in Kapitel 1 detailliert beschrieben wird, liegt eine weitere Fehlerquelle in der Programmierung von Onlinefragebögen. Aus sorgloser Programmierung können Darstellungs- und Funktionalitätsunterschiede von Fragebögen resultieren, was im äußersten Fall zu selektiven Teilnahmeabbrüchen führen kann. Gerade bei Antwortskalen können geringe Änderungen des Aussehens einen großen Einfluss auf die Datengüte haben (siehe Couper, Conrad, & Tourangeau, 2007; Symth, Christian, Dillman, & Stern, 2006b). Der online forschende Wissenschaftler benötigt somit zusätzliches Wissen, damit er aus unzureichender Programmierung resultierende Probleme erkennen und vermeiden kann. Erfolgreiche Forschung im Internet bedarf mehr, als einen vorhandenen Fragebogen in ein Textverarbeitungsprogramm zu kopieren, ihn als Website zu speichern und online zugänglich zu machen.

i.i.ii Effekte von Antwortskalen

Antwortskalen, beispielsweise zur Messung des Grades der Zustimmung zu einem Testitem, spielen bei der Durchführung von Untersuchungen eine besondere Rolle, da über sie ein wichtiger Teil der Kommunikation zwischen Forscher und Untersuchungsteilnehmer erfolgt (siehe Krosnick & Fabrigar, 1997). Bei geschlossenen Ratingskalen kann eine größere Anzahl von Antwortmöglichkeiten beispielsweise anzeigen, dass eine elaborierte Antwort erwartet wird, und eine geringe Anzahl von Antwortmöglichkeiten, dass lediglich ein globales Urteil gefragt ist. Vor allem bei unklarer Fragestellung oder Konzeptualisierung nutzen Untersuchungsteilnehmer verstärkt Eigenschaften von Antwortskalen, um Rückschlüsse auf die Be-

deutung einer Frage zu ziehen und greifen dafür auf bestimmte Heuristiken zurück (z. B. dass die mittlere Antwortmöglichkeit als typische Antwort angesehen wird; siehe Tourangeau, Couper, & Conrad, 2004).

Neben den oben genannten eher allgemeinen Fehlerquellen bergen Antwortskalen ein besonderes Fehlerpotenzial, unter anderem deswegen, da bereits geringe Darstellungsunterschiede einen bedeutsamen Einfluss auf das Antwortverhalten haben können. Dies ist in der allgemeinen Umfragemethodologie gut dokumentiert (Lyberg et al., 1997; Schwarz, Hippler, Deutsch, & Strack, 1985; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991) und trifft auch auf Webbasierte Befragungen zu (siehe Couper, 2008; Couper, Traugott, & Lamias, 2001; Dillman et al. 2009; Dillman & Bowker, 2001; Reips, 2010; Smyth, Dillman, Christian, & Stern, 2006; Tourangeau, Couper, & Conrad, 2004). In Kapitel 1 werden mögliche Probleme verschiedener Antwortskalen im Detail diskutiert.

i.i.iii Visuelle Analogskalen

Ein besonderer Skalentyp, der bereits in den 1920er Jahren beschrieben wurde (Hayes & Patterson, 1921), sind visuelle Analogskalen (VAS, siehe auch Kapitel 1.3.6). Im Großteil der hier vorgestellten Studien (Kapitel 2, 3 und 4) wird der Einfluss von VAS auf die Güte der damit erhobenen Daten untersucht. Bei VAS handelt es sich um ein kontinuierliches Messinstrument, bei dem Befragte das Ausmaß einer Antwort (z. B. der Zustimmung mit einer Aussage) auf einem graphisch repräsentierten Kontinuum – einer Linie – verorten. Eine feste Operationalisierung von VAS findet sich in der Literatur nicht, sodass die konkrete Ausgestaltung dem Forscher obliegt. Bei den im Rahmen dieser Arbeit genutzten VAS waren ausschließlich die Enden mit Ankerstimuli versehen und jeder Pixel in der Länge der VAS einer möglichen Antwort. Abbildung i.i verdeutlicht, wie ein Urteil auf VAS abgegeben wird: Im Ruhezustand, direkt nach dem Laden der Webseite, ist die VAS ohne Markierung (A). Zunächst muss der Mauszeiger an die Stelle bewegt werden, die den zutreffenden Wert repräsentiert (B). Mit einem Klick wird die Antwort gegeben, die durch ein Kreuz auf der Skala kenntlich gemacht wird (C). Das Urteil kann durch einen Klick auf eine andere Stelle der VAS korrigiert werden.

VAS beeinflussen drei Stadien empirischer Studien: die Fragebogenentwicklung, den Messprozess und die Datenanalyse. Bei der *Fragebogenentwicklung* entfällt die bei diskreten Antwortskalen nötige Entscheidung für eine gewisse (gerade oder ungerade) Anzahl von Antwortmöglichkeiten; mögliche Kategorisierungen erfolgen erst während der Analyse.

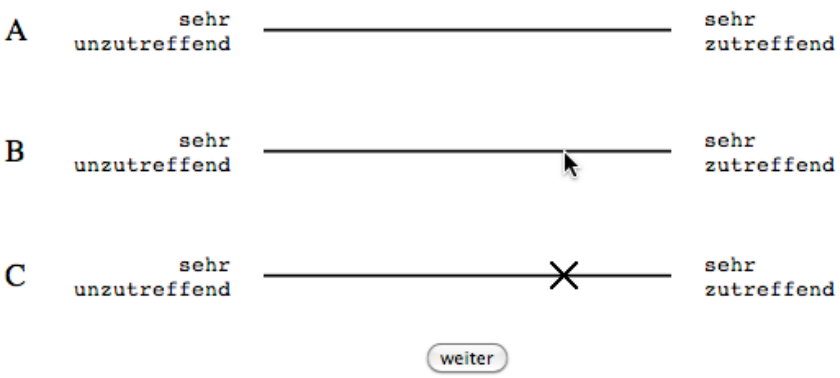


Abbildung i.i. VAS mit 250 möglichen Werten direkt nach dem Laden einer Webseite (A), mit Mauszeiger über dem zutreffenden Wert (B) und mit durch Klick abgegebenem Urteil (C).

Für den *Messprozess* bieten VAS viele Vorteile. Der offenkundigste Vorteil ist, dass schon geringste Unterschiede oder Veränderungen im Antwortverhalten gemessen werden können. Zudem sind VAS sehr effizient hinsichtlich der Nutzung des auf einem Bildschirm verfügbaren Raumes. Eine VAS mit einer Spannweite von 50 möglichen Messpunkten nimmt denselben Platz ein wie drei Radiobuttons. Dass Messungen auch mit sehr kurzen VAS valide sein können, zeigt die in Kapitel 2 vorgestellte Untersuchung (Reips & Funke, 2008). Der testtheoretisch wichtigste Punkt ist allerdings, dass mit VAS erhobene Daten als intervallskaliert betrachtet werden können (siehe Kapitel 2). Bei Kategorialskalen mit einer begrenzten Anzahl von Antwortmöglichkeiten gibt es Hinweise darauf, dass Befragte den Antwortkategorien (vor allem den Randkategorien, siehe Funke, 2004; Funke & Reips, 2006) unterschiedlich große Intensitätsintervalle zuweisen. Während bei Messungen mit Kategorialskalen also fraglich ist, ob Daten intervallskaliert sind, erfüllen mit VAS erhobene Daten dieses Kriterium für die Anwendung einer Vielzahl statistischer Testverfahren.

Bezüglich der *Datenanalyse* erweitern VAS die Möglichkeiten erheblich. Die Rohdaten – also die unmittelbar auf einer VAS gegebenen Antworten – können zunächst unverändert genutzt werden. Mit den Rohdaten kann im Vergleich zu Kategorialskalen auch präziser getestet werden, ob Daten einer bestimmten Verteilung folgen – normalverteilte Daten sind beispielsweise eine Voraussetzung für einige statistische Verfahren. Das eigentlich Besondere bei der Datenauswertung mit VAS ist allerdings, dass im Rahmen des Analyseprozesses *parallel* mehrere unterschiedliche Kategorisierungen (z. B. Dichtomisierung oder Gruppierung in Quantile) durchgeführt werden können, sodass mit ein und demselben Datensatz mehrere Analysen berechnet werden können. Im Gegensatz dazu sind die Möglichkeiten mit Kategorialskalen sehr begrenzt. So kann beispielsweise eine Dichotomisierung mit gleichviel Kategorien umfassenden Gruppen nur bei Antwortskalen mit einer geraden Anzahl von Kategorien erfolgen.

Visuelle Analogskalen in der akademischen Forschung. Bedenkt man die obigen Vorteile, die bei Messungen mit VAS zu erwarten sind, sollte man meinen, dass diese Antwortskala zu den am häufigsten genutzten Messinstrumenten gehört. Eine Sichtung der Fachliteratur zeigt allerdings, dass VAS in der Psychologie ebenso wie in der Soziologie und angrenzenden Wissenschaften, in denen viel mit Fragebögen gearbeitet wird, kaum eingesetzt werden. Einzig in der Medizin – vor allem in der Schmerzforschung – werden VAS häufig genutzt (siehe Couper et al., 2006). Für diese geringe Verbreitung von VAS können neben einer generellen Innovationsaversion mehrere Faktoren als ursächlich angesehen werden: Erwägungen hinsichtlich des praktischen Aufwands und der technischen Umsetzung, eine unzureichende Berücksichtigung des Messfehlers und das Bestreben nach Forschungskontinuität. Wie im Folgenden verdeutlicht wird, spricht bei der computergestützten Datenerhebung lediglich in speziellen Situationen der letzte Punkt gegen die Nutzung von VAS.

Die wichtigste Erklärung für die seltene Nutzung von VAS liegt wahrscheinlich darin, dass bei der Entscheidung für eine bestimmte Antwortskala deren Einfluss auf den Messfehler nur unzureichend berücksichtigt wird. Grundlegend wird zwischen systematischer Antwortverzerrung (*Bias*) und zufälliger Abweichung (*Error*) unterschieden. Im ersten Fall tritt eine Über- oder Unterschätzung der zentralen Tendenz von Daten (wie arithmetisches Mittel und Median) auf. Bei einem Großteil der in der Literatur berichteten, direkten Vergleichen von Messungen mit VAS und Kategorialskalen sind Unterschiede zwischen den Skalen entweder statistisch oder praktisch nicht signifikant (siehe Couper et al., 2006 für einen webbasierten Vergleich). Aus diesen Ergebnissen wird häufig fälschlicherweise abgeleitet, dass VAS keinen Einfluss auf den Messprozess haben und folglich der einfacher zu administrierenden Skala, also der Kategorialskala, der Vorzug zu geben ist. Das Fehlen einer systematischen Verzerrung ist allerdings kein hinreichendes Kriterium äquivalenter Messungen. Der Schlüssel zum tieferen Verständnis der Skalengüte liegt in der Betrachtung des zufälligen Messfehlers. Der zufällige Messfehler – auch das Rauschen in den Daten genannt – beeinflusst zwar nicht die Messung von Mittelwerten und führt zu erwartungstreuen Schätzern der zentralen Tendenz, er hat jedoch andere negative Auswirkungen. Das Datenrauschen erhöht den Standardfehler von Messungen, was zu einer *generellen Unterschätzung von Effektstärken* führt. Da der Standardfehler auch die Irrtumswahrscheinlichkeit statistischer Tests beeinflusst, wird zudem die Nullhypothese in unzutreffender Weise aufrechterhalten, der Fehler 2. Art steigt und die Teststärke (*Power*) sinkt. Um tatsächlich bestehende Unterschiede auf gegebenem statistischem Signifikanzniveau messen zu können, muss die Anzahl der Untersuchungsteilnehmer erhöht werden. *Kleine Effekte* bleiben allerdings auch bei einer Erhöhung der Fallzahl unbeobachtbar. Kleine, aber inhaltlich bedeutsame Unterschiede können also

nur gemessen werden, wenn das Rauschen in den Daten möglichst gering ist. In Kapitel 3 wird der positive Einfluss aufgezeigt, den VAS im Vergleich zu Kategorialskalen auf den zufälligen Messfehler haben.

Vor allem bei großen Untersuchungen spielen finanzielle Erwägungen bei der Fragebogengestaltung eine wichtige Rolle. Bei Papierfragebögen bedeutet die Dateneingabe bei VAS einen erheblichen Mehraufwand im Vergleich zu Kategorialskalen, da die exakte Position jeder Antwortmarkierung aufwändig ausgelesen werden muss (siehe Funke, 2004). Dieser Nachteil entfällt bei computerisierten Untersuchungen aber vollständig, da das Auslesen schnell und fehlerfrei geschieht. Doch selbst bei der Forschung im Web werden VAS kaum genutzt. Dies mag darin begründet sein, dass sich VAS nicht mit HTML, der grundlegenden Technik bei der Entwicklung von Fragebögen, erzeugen lassen. Auch in professioneller Umfrage- und Testsoftware sind VAS kaum verbreitet, sodass das Einbinden von VAS in Onlinestudien häufig mit einem erheblichen Mehraufwand verbunden ist. Mit dem in Kapitel 1 vorgestellten *VAS Generator* kann diese technische Hürde problemlos genommen werden, sodass VAS auch ohne spezielle Kenntnisse in Onlinebefragungen genutzt werden können.

Problematisch ist der Einsatz von VAS lediglich dann, wenn Daten aus unterschiedlichen Studien möglichst gut vergleichbar sein sollen. Wie oben bereits beschrieben, können Antwortskalen das Antwortverhalten stark beeinflussen. Um eine größtmögliche Vergleichbarkeit von Ergebnissen – entweder im Längschnitt oder zwischen unabhängigen Untersuchungen – gewährleisten zu können, sollten daher Änderungen des Messinstruments vermieden werden. Für die Entwicklung neuer Messinstrumente sollte aber die bestmögliche Antwortskala genutzt werden. Ob es sich dabei um eine Kategorialskala mit einer bestimmten Anzahl von Antwortoptionen oder um eine VAS handelt, hängt auch von den konkreten Zielsetzungen einer Studie ab. Grundsätzlich ist immer dann zu VAS zu raten, wenn *kontinuierliche* latente Variablen gemessen werden sollen.

Die oben genannten Punkte bieten eine Erklärung dafür, warum vor allem im medizinischen Bereich mit VAS gearbeitet wird. So werden medizinische Untersuchungen häufig an relativ kleinen Stichproben durchgeführt und gerade bei kontinuierlichen, subjektiven Phänomenen wie Schmerz sind bereits kleinste Veränderungen für Betroffene von großer Bedeutung. Dazu kommt, dass im Vergleich zu den Kosten für die Entwicklung einer neuen Therapie der erhöhte Aufwand für das Auslesen von VAS nicht ins Gewicht fällt. Letztlich werden VAS auch aus einer gewissen Fachkultur heraus genutzt und um verschiedene Studien möglichst gut miteinander vergleichen zu können.

i.ii Zielsetzung

In der vorliegenden Arbeit wurde geprüft, ob die allgemeine Zurückhaltung in der Anwendung außerhalb der Medizin trotz der zahlreichen für die Nutzung von VAS sprechenden Gründe gerechtfertigt ist. Das Hauptanliegen dieser Dissertation ist, einen Beitrag dazu zu leisten, die erhebliche Forschungslücke zu webbasierten VAS zu schließen. Die Ergebnisse der vorgestellten Untersuchungen sollen Forschern die Entscheidung für oder gegen den Einsatz von VAS in computergestützten Untersuchungen erleichtern.

In fünf Kapiteln wird das Potenzial graphischer Antwortskalen – vor allem von VAS – untersucht. In *Kapitel 1* (Funke & Reips, 2007) werden die Grundlagen und Möglichkeiten von Messungen im Web dargestellt. Anhand relevanter Befunde aus der umfragemethodologischen Literatur wird aufgezeigt, welchen Einfluss unterschiedliche Antwortskalen (u. a. Kategorialskalen, VAS und Sliderskalen) auf die gegebenen Antworten haben können. In *Kapitel 2* (Reips & Funke, 2008) wird der VAS Generator vorgestellt, mit dem sich online kostenfrei VAS erstellen lassen. In Übereinstimmung mit den theoretischen Annahmen ihrer Messeigenschaften wird gezeigt, dass mit VAS erhobene Daten als intervallskaliert behandelt werden können. In *Kapitel 3* (Funke & Reips, 2010a) wird zunächst eine Simulationsstudie dargestellt, die den Einfluss der Anzahl der Antwortmöglichkeiten bei Kategorialskalen auf den Messfehler untersucht. In zwei empirischen Folgeuntersuchungen wird wie vorhergesagt beobachtet, dass VAS das Rauschen in den Daten verringern und einen praktisch bedeutsamen positiven Effekt auf die Datengüte haben. *Kapitel 4* (Funke & Reips, 2010b) hat eine Untersuchung zum Thema, die zeigt, dass sich Messungen mit semantischen Differentiale durch die Nutzung von VAS optimieren lassen. Zum Abschluss werden in *Kapitel 5* (Funke, Reips, & Thomas, in press) schwerpunktmäßig Sliderskalen untersucht. Dabei handelt es sich um einen Skalentyp, der VAS zwar äußerlich ähnelt (zur Abgrenzung siehe Kapitel 1), sich aber in wichtigen Charakteristika – wie Datenqualität und Art der Nutzung – substantiell von ihnen unterscheidet. Sliderskalen erwiesen sich in dem vorgestellten Webexperiment als ein problematisches Messinstrument, da vor allem Untersuchungsteilnehmer mit einer geringen formalen Bildung verstärkt Probleme mit deren Nutzung haben.

i.iii Zusammenfassung der zentralen Ergebnisse

i.iii.i Kapitel 1: Web-based measurement and rating scales

Findet Forschung nicht in einem abgeschlossenen und gut kontrollierbaren System wie einem Labor statt, sondern in einem offenen und schwer kontrollierbaren Raum wie dem Internet, tun sich neben neuen Möglichkeiten auch neue Probleme auf. In diesem Kapitel (basierend auf Funke & Reips, 2007) wird grundlegend in das Thema Onlineforschung eingeführt und eine Übersicht über die Möglichkeiten und Fallstricke Internet-basierter Forschung gegeben.

Vor allem technische Variablen und die damit verbundene Gefahr der Antwortverzerrung (siehe Buchanan & Reips, 2001) sind online von Bedeutung. Während sich im Labor vor Ort die technische Funktion einer am Computer durchgeführten Studie relativ einfach sicherstellen lässt, kann bei online durchgeführten Studien die technische Varianz seitens der Untersuchungsteilnehmer zu Problemen führen. So können Software wie der Webbrowser und Hardware wie der Bildschirm die Darstellung und Funktionalität eines Onlinefragebogens beeinträchtigen und einen Einfluss auf Teilnahmebereitschaft und Datengüte haben (siehe Dillman et al., 2009; Reips, 2002b, 2002c; Schmidt, 2007). Während negative Effekte (wie Teilnahmeabbrüche und Antwortverzerrungen) bei sorglos gestalteten Untersuchungen fast schon zwangsläufig sind, bieten methodisch sauber umgesetzte Designs einen Mehrwert im Vergleich zu im Labor durchgeführten Studien. Zu den besonderen Möglichkeiten zählt, in kurzer Zeit – wie bereits oben angesprochen – Studien mit einer großen Anzahl geographisch weit von einander entfernter Untersuchungsteilnehmer durchzuführen und vor allem die Anreicherung von Untersuchungsergebnissen mit nicht-reaktiv erhobenen Prozessdaten wie beispielsweise Reaktionszeiten.

Forscher, die ihr Labor auf das Internet ausdehnen möchten, müssen sich zwangsläufig auch mit technischen Belangen beschäftigen, um die Vorteile der Onlineforschung nutzen und die Nachteile vermeiden zu können. Dabei sollte bereits in der Entwicklung eines Untersuchungsdesigns beachtet werden, welche Möglichkeiten und welche potentiellen Fehlerquellen mit welcher technischen Umsetzung verbunden sind. Es gilt, zwischen allgemeinen Effekten und Möglichkeiten (z. B. Filterführung, Personalisierung, Randomisierung) computergestützter Fragebögen und der speziellen Situation bei Onlinebefragungen (z. B. Nutzung von Prozessdaten, Minimierung technischer Varianz) zu unterscheiden. Forschung im Internet kann nur dann erfolgreich sein, wenn sich das vorhandene methodologische Wissen in einem adäquaten Forschungsdesign widerspiegelt. Die hier zusammengetragenen Befunde

der Umfrageliteratur zeigen deutlich, dass der unreflektierte Gebrauch technischer Möglichkeiten ein großes Risikopotenzial birgt und es bei Untersuchungen im Web sinnvoll ist, robuste *Lowtechverfahren* – bei denen die technischen Anforderungen an den Computer des Teilnehmers so gering wie möglich sind – zu wählen (siehe auch Schwarz & Reips, 2001).

i.iii.ii Kapitel 2: Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator

In Kapitel 2 (Reips & Funke, 2008) wird der kostenfreie Webservice *VAS Generator* (http://vasgenerator.net) vorgestellt, der im Rahmen dieser Dissertation erstellt wurde und im August 2006 online ging. Mit VAS Generator (siehe Abbildung 2.2, Seite 50) lassen sich in fünf Schritten VAS für die browserbasierte Forschung erstellen. Es können Länge, Breite und Farbe von VAS, des genutzten Markers sowie die Endanker variiert werden. Nach einem Test der individuell erstellten Antwortskala lassen sich alle benötigten Dateien herunterladen und für die Forschung im Web oder im lokalen Labor nutzen. Die Website wird rege genutzt und wurde insgesamt bereits über 17 000-mal besucht[i.1], allein im ersten Halbjahr 2010 über 5 000-mal (siehe Abbildung i.ii).

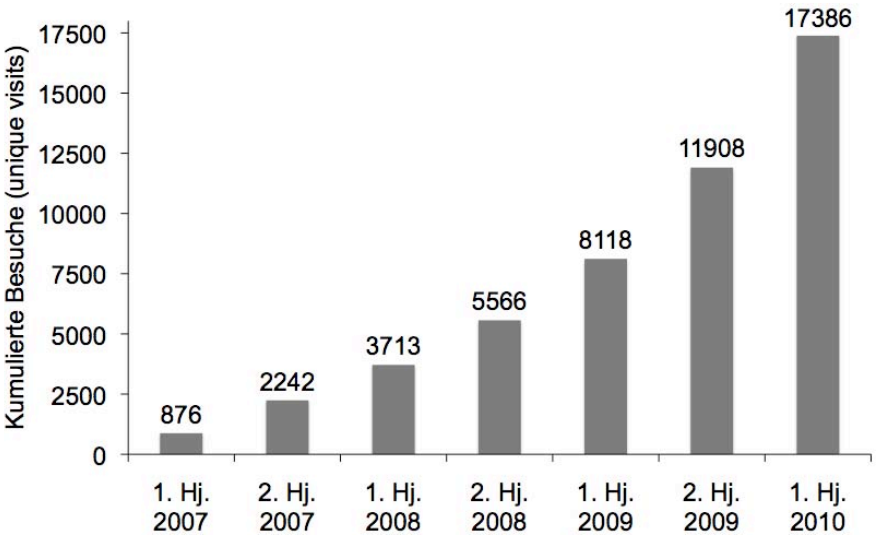


Abbildung i.ii. Kumulierte Anzahl der Besuche auf http://vasgenerator.net zwischen dem 1. Halbjahr (Hj.) 2007 und dem 1. Halbjahr 2010.

[i.1] Als ein Besuch (*unique visit*) werden alle Zugriffe mit derselben IP-Adresse und Browserkennung gewertet, zwischen denen mehr als 30 Minuten liegen.

Um den Einfluss der mit VAS Generator erstellten Antwortskalen auf die Datengüte zu testen, wurde ein Webexperiment durchgeführt. In einen 2 × 3 Design wurde die Präzision gemessen, mit der Untersuchungsteilnehmer 13 vorgegebene Prozentwerte oder Bruchzahlen auf mit VAS Generator erstellten Skalen platzierten. Als zweiter Faktor wurde die Länge der VAS manipuliert (50 Pixel, 200 Pixel oder 800 Pixel), um zu testen, wie robust Messungen mit VAS bei Darstellungsunterschieden aufgrund technischer Varianz sind.

Der Hauptbefund des geschilderten Experiments ist, dass mental gut repräsentierte Werte wie Zahlen mit großer Präzision (die mittlere Abweichung lag bei drei Prozentpunkten) auf VAS verorten werden können. Selbst bei ausgesprochen kurzen und langen VAS entsprachen gleichgroße numerische Intervalle gleichgroßen Abschnitten auf der VAS. Dies ist ein deutlicher Beleg dafür, dass mit VAS erhobene Daten äquidistant sind und intervallskaliert behandelt werden können (siehe auch Hofmann & Theuns, 2008; Myles, Troedel, Boquest, & Reeves, 1999; Myles & Urquhart, 2005). Somit sind die testtheoretischen Voraussetzungen für die Anwendung einer großen Zahl statistischer Verfahren gegeben.

i.iii.iii Kapitel 3: Lowering measurement error with visual analogue scales

In diesem Kapitel (Funke & Reips, 2010a) wird untersucht, in welchem Ausmaß VAS und verschiedene Kategorialskalen den Messfehler beeinflussen (zum Konzept des *Total Survey Error* siehe Groves et al., 2004). Im Zentrum der durchgeführten Untersuchungen steht der Formatierungsfehler, der vor allem dann auftritt, wenn kontinuierliche Variablen mit Kategorialskalen gemessen werden. Der Formatierungsfehler ist die Abweichung zwischen dem Wert, der auf einer Antwortskala berichtet wird, und dem tatsächlich zutreffenden Wert (siehe Schwarz & Oyserman, 2001). Abbildung i.iii illustriert die Entstehung dieser Fehlerquelle. Bei diskreten Ratingsskalen hängt der Formatierungsfehler eng mit der Anzahl der Antwortoptionen zusammen.

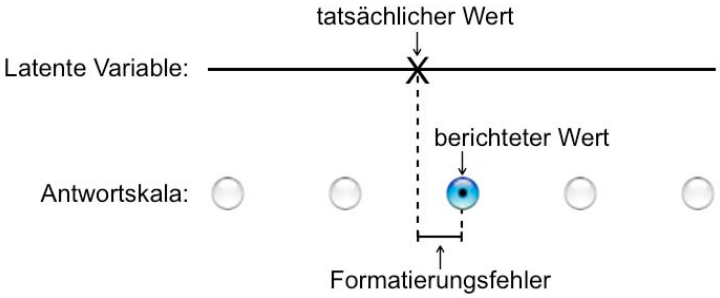


Abbildung i.iii. Illustration des Formatierungsfehlers bei einer 5-stufigen Kategorialskala.

Bei der Planung einer Untersuchung mit Kategorialskalen stellt sich dem Forscher die Frage, wie viele Antwortoptionen den Untersuchungsteilnehmern (beispielsweise für die Bewertung eines Items in einem Persönlichkeitstest) zur Verfügung gestellt werden. Für oder gegen eine gewisse (gerade oder ungerade) Anzahl von Antwortmöglichkeiten sprechen neben der Art des Items (unipolar versus bipolar) und der (individuell und nach Item unterschiedlichen) Diskriminationsfähigkeit der Befragten auch messtheoretische Überlegungen eine Rolle. Darüber hinaus hat die Anzahl der Antwortkategorien auch einen Einfluss auf die Reliabilität und Validität von Messungen (siehe Lozano, García-Cueto, & Muñiz, 2008).

Um die Wahl für eine Antwortskala anhand des damit verbundenen Messfehlers treffen zu können, wurden drei unabhängige Untersuchungen durchgeführt. In Studie 1, einer grundlegenden Simulationsstudie, wurde der erwartete Formatierungsfehler sowohl bei Einzelmessungen, als auch bei Aggregatdaten bei VAS und bei Kategorialskalen mit einer unterschiedlichen Anzahl von Antwortmöglichkeiten berechnet. Neben dem – wenig erstaunlichen – Befund, dass der Formatierungsfehler bei einer Einzelmessung mit steigender Kategorienanzahl geringer wird, ermöglicht die Studie eine *Quantifizierung* des minimal zu erwartenden Formatierungsfehlers. Hauptbefund bei der Analyse aggregierter Daten ist, dass es eine Interaktion zwischen der Verteilung der tatsächlichen Werte und der Anzahl der Antwortkategorien gibt. Eine größere Anzahl von Antwortkategorien führt also nicht in jedem Fall zu einer Verringerung des Formatierungsfehlers. In den Studien 2 und 3 wurde der theoretisch hergeleitete positive Einfluss von VAS auf die Datengüte empirisch bestätigt. Der skalenbedingte Formatierungsfehler (operationalisiert durch den Standardfehler des Mittelwerts) ist mit VAS im Vergleich zu 5- und 7-stufigen Antwortskalen geringer. Erst beim Vergleich mit 9-stufigen Skalen ließ sich kein Effekt mehr beobachten.

Die aus dieser Untersuchung resultierende Empfehlung lautet eindeutig, VAS als Antwortskala zu nutzen, da sie zu geringeren Messfehlern führen. Zum einen bedeutet ein geringerer Messfehler, dass bereits sehr kleine Effekte gemessen werden können, die ansonsten im Rauschen der Daten untergehen würden. Zum anderen lassen sich größere Effekte bereits mit einer geringeren Anzahl von Untersuchungsteilnehmern messen, was die Untersuchungskosten verringert. Nur falls methodische Argumente wie das Lowtechprinzip gegen den Einsatz von VAS sprechen oder die für die Darstellung von VAS erforderliche Technik bei einem Teil der Befragten nicht verfügbar ist, sollte auf Kategorialskalen zurückgegriffen werden.

i.iii.iv Kapitel 4: Semantic differentials made of visual analogue scales

Ein spezielles Einsatzgebiet für VAS wird in Kapitel 4 untersucht (Funke & Reips, 2010b). Bei semantischen Differentialen werden verschiedene Dimensionen ein und derselben Variable gleichzeitig und in Relation zu einander erhoben (siehe Osgood, 1952; Osgood, Suci, & Tannenbaum, 1957). VAS scheinen als Antwortskala in semantischen Differentialen besonders gut geeignet zu sein, da mit diesem Skalentyp feine Unterschiede zwischen den einzelnen Dimensionen kommuniziert und somit ein stimmiges Gesamtbild gezeichnet werden kann. Im Gegensatz dazu ist die Kombination von semantischen Differentialen und Kategorialskalen deutlich problematischer. Abbildung i.iv zeigt semantische Differentiale mit VAS und mit Kategorialskalen. Das Abgeben eines konsistenten Urteils hinsichtlich aller Dimensionen wird bei Kategorialskalen bereits dadurch erschwert, dass unvermeidlich das Problem geteilter Ränge auftritt, wenn die Anzahl der zu bewertenden Dimensionen die Anzahl der Antwortmöglichkeiten überschreitet. Daraus resultiert das Artefakt, dass einzelne Dimensionen als gleichwertig angegeben werden, obwohl tatsächlich Unterschiede in ihrer Wertigkeit vorliegen.

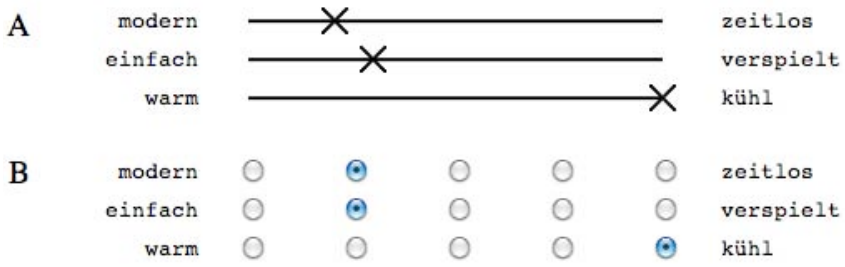


Abbildung i.iv. Semantisches Differential mit VAS (A) und Kategorialskala (B).

Während des vierten im Rahmen dieser Arbeit durchgeführten Webexperiments beantworteten Untersuchungsteilnehmer semantische Differentiale, die entweder aus 5-stufigen Kategorialskalen oder aus VAS bestanden. Als Resultat ergaben sich deutliche Unterschiede im Antwortverhalten. Bei gleicher Antwortdauer korrigierten die Befragten ihre Urteile mit VAS häufiger, als mit Kategorialskalen. Die Analyse der gegebenen Antworten, der Häufigkeit der Antwortkorrektur und der Antwortdauer zeigt, dass mit VAS die Präzision des Urteils im Vergleich zu Kategorialskalen maximiert wird. Insgesamt betrachtet erscheint es sehr sinnvoll, bei semantischen Differentialen VAS anstelle von Kategorialskalen als Antwortskalen einzusetzen, da mit VAS ein konsistenteres Urteil hinsichtlich aller erfragten Aspekte gegeben werden kann.

i.iii.v Kapitel 5: Sliders for the smart: Type of rating scale on the Web interacts with educational level

Im abschließenden Teil der vorliegenden Arbeit werden nicht VAS sondern Sliderskalen untersucht (Funke, Reips, & Thomas, in press). Auf den ersten Blick handelt es sich um VAS sehr ähnliche graphische Antwortskalen (siehe Abbildung i.v). Sliderskalen sind im Vergleich zu den kontinuierlichen VAS aber diskrete Messinstrumente, bei denen lediglich eine geringe Anzahl von Antwortmöglichkeiten zur Auswahl steht. In der Art der Benutzung liegt ein weiterer wichtiger Unterschied. Urteile werden bei Sliderskalen mit einem Schieberegels abgegeben, der angeklickt und verschoben werden muss, was ein größeres Maß an Hand-Auge-Koordination erfordert. In der vorgestellten Studie wurden zwei Faktoren experimentell manipuliert, die Antwortskala (Sliderskala oder Kategorialskala) und die (horizontale oder vertikale) Ausrichtung der Skala auf dem Bildschirm.

Abbildung i.v. Sliderskala mit Schieberegeler in mittlerer Position.

Hauptbefund des Webexperimentes ist, dass die hohen technischen Anforderungen von Sliderskalen nicht nur dazu führen, dass die Abbruchquote generell steigt, sondern dass es einen starken Zusammenhang zwischen der formalen Bildung der Befragten und dem Auftreten von Problemen mit Sliderskalen gibt. Personen mit geringer Schulbildung brachen die Teilnahme an der Untersuchung häufiger ab, da sie offensichtlich mit Sliderskalen überfordert waren. Dieser selektive Abbruch kann ernsthafte Antwortverzerrungen zur Folge haben (siehe Buchanan & Reips, 2001). Darüber hinaus erweisen sich die getesteten Sliderskalen als ausgesprochen negativ hinsichtlich der Antwortverteilung und der Antwortdauer.

Beim Experimentalfaktor Drehung der Antwortskala wurde bei keiner der beiden Ratingskalen ein bedeutsamer Effekt hinsichtlich zentraler Tendenz, Abbruchquote und Verteilung der Werte beobachtet. Insgesamt sind dies Indizien dafür, dass Änderungen in der Ausrichtung von Antwortskalen als Designelement genutzt werden können. Ein möglicher Anwendungsbereich ist das (automatische) Anpassen der Antwortskala an das Format des vorhandenen Bildschirms. Dies kann vor allem bei mobilen Onlinebefragungen, bei denen das Endgerät zum Beispiel ein Mobiltelefon (deren Display meist hochformatig) ist, hilfreich sein, da somit alle Antwortmöglichkeiten direkt sichtbar sind (siehe Couper, Tourangeau, Conrad, & Crawford, 2004).

Insgesamt bestätigt sich in dieser Untersuchung, dass Antwortskalen im Allgemeinen so einfach wie möglich gestaltet werden sollten (siehe auch Buchanan & Reips, 2001; Reips, 2002c). Sliderskalen erweisen sich im Vergleich zu Kategorialskalen als unvorteilhaft, da ihre Nutzung mit erhöhten kognitiven Anstrengungen verbunden ist. Zusätzliche Probleme sind mit dem Schieberegler verbunden. Seine Ausgangsposition (z. B. in der Mitte der Skala) kann sowohl die Identifizierung von Antwortverweigerung erschweren, als auch zu Ankereffekten und somit zu Antwortverzerrung führen. Ein Vergleich zwischen VAS, Kategorialskalen und Sliderskalen hat zudem gezeigt, dass Sliderskalen und Kategorialskalen zu sehr ähnlichen Daten führen, die beide deutlich von mit VAS erhobenen Daten abweichen (vgl. Funke & Reips, 2008, siehe auch Appendix B). Es ist also nicht ausreichend, einfach eine Linie hinter eine diskrete Antwortskala zu zeichnen, wie es bei Sliderskalen der Fall ist. Die Messung kontinuierlicher latenter Variablen bedarf auch eines angemessenen, kontinuierlichen Messinstruments.

i.iv Schlussfolgerungen

Zusammenfassend wird aus den vorgestellten Studien zu VAS und Sliderskalen deutlich, dass Antwortskalen einen bedeutenden Einfluss auf die Quantität und Qualität online erhobener Daten haben können, auch wenn Antwortskalen nur einen gewissen Teil des Messprozesses beeinflussen (siehe Schwarz & Oysermann, 2001; Groves et al., 2009). Mit einem auf anderen Ebenen – zum Beispiel bei der Frageformulierung oder der graphischen Gestaltung – schlecht entwickelten Fragebogen können auch mit der besten Antwortskala keine validen Daten erhoben werden. Eine Skala mit schlechten Messeigenschaften kann allerdings selbst bei einer ansonsten wohldurchdachten Untersuchung die Datenqualität deutlich verringern. So sollten Antwortskalen bei der Entwicklung eines Fragebogens sorgfältig ausgewählt werden und nicht losgelöst von Befragungsinhalt, Befragtencharakteristika und technischen Möglichkeiten gesehen werden. Insbesondere die in Kapitel 5 dargestellte Untersuchung zu Sliderskalen zeigt, dass ein sorgloser Umgang mit Technik und Antwortskalen einen desaströsen Einfluss auf die Datenerhebung haben kann. Die Untersuchungen zu VAS machen jedoch auch deutlich, dass Antwortskalen, die von den etablierten Kategorialskalen abweichen, durchaus sehr positive Effekte haben können. Wie in den Kapiteln 2 bis 4 dargestellt, sind VAS eine äußerst wertvolle Erweiterung des Methodenspektrums der Onlineforschung.

Die vorliegende Dissertation hat eine erkenntnistheoretische und eine ethische Dimension. Zum einen steht in allen Kapiteln mehr oder weniger explizit die Frage zur Disposition, wel-

chen Einfluss die verwendete Antwortskala auf den *Erkenntnisgewinn* hat. Um kleine – trotz ihrer geringen absoluten Größe inhaltlich bedeutsame – Unterschiede messen zu können, muss das Fehlerrauschen in den Daten so weit wie möglich verringert werden. Die Untersuchungen zu VAS haben gezeigt, dass die Verringerung des Messfehlers nicht nur eine theoretische Eigenschaft der Skala ist, sondern auch eine empirische Entsprechung hat. Somit ist der Einsatz von VAS in der Onlineforschung dem Erkenntnisgewinn zuträglich.

Psychologische Forschung hängt zu einem großen Teil von der Kooperation der Untersuchungsteilnehmer ab. Es sollte nicht nur aus ökonomischem Kalkül Gewicht auf eine möglichst geringe Teilnehmerzahl gelegt werden. Ein Aspekt *ethisch verantwortungsvollen* Forschens besteht darin, der Antwortskala Vorzug zu geben, die die benötigte Teilnehmerzahl verringert. Wie die in Kapitel 3 beschriebene Reduzierung des Messfehlers zeigt, können VAS dazu einen Betrag leisten und sollten somit auch unter dem Gesichtspunkt guter wissenschaftlicher Praxis genutzt werden.

Insgesamt sind die hier vorgestellten methodologischen Untersuchungen außerordentlich vielversprechend. VAS können Messungen mental gut repräsentierter, kontinuierlicher latenter Variablen deutlich verbessern. Die kaum verbreitete Nutzung von VAS in der psychologischen Forschung entbehrt nach den hier gewonnenen Erkenntnissen zumindest bei Untersuchungen im Internet jeglicher Grundlage. An hoher Datenqualität interessierte Forscher sollten das Potenzial visueller Analogskalen für ihre webbasierten Studien ausschöpfen.

1

Web-Based Measurement and Rating Scales

1 Web-Based Measurement and Rating Scales[1.1]

1.1 Introduction

The widespread access to the Internet makes the World Wide Web an interesting medium for many areas of science, especially for psychology. Computer assisted Web interviewing (CAWI) and other reactive or non-reactive Web-based methods broaden the possibilities for qualitative and quantitative research. One reason for their increasing popularity is the fast access to large samples and rare populations (like for instance people suffering from sex-somnia, see Mangan & Reips, 2007). More and more software and Web services – often freely available – make it possible even for researchers without technical background to collect data online. Web laboratories and Web sites listing studies facilitate an easy recruitment of respondents from all over the world (e.g., Reips, 2001; Reips & Lengler, 2005). It is sometimes disregarded, though, that *quantity* is only one aspect whereas high data *quality* is the fundamental basis for valid conclusions.

Internet-based data collection can eliminate some problems of paper-based questionnaires (e.g., presentation of stimuli that cannot be printed or complex filtering). Some other problems arise in both modes (like social desirability bias or visual design effects), and there are also some new challenges (e.g., browser compatibility, data generated by automatic scripts instead of human respondents, and multiple participation). But most importantly, new opportunities open up that cannot be realized – or only with considerable effort – in paper and pencil studies like the availability of very large samples, the use of multimedia stimuli (Fuchs & Funke, 2007), and truly voluntary participation. Moreover, computerized questionnaires allow the unobtrusive, non-reactive collection of paradata (see Heerwegh, 2003) like response time, which is especially valuable if each item is presented on a separate Web page (see Reips 2002b).

The design of a study influences both quantity and quality of data (for an overview see Couper, 2008). The two main influences regarding data quality (for an overview see Groves et al., 2009) are representativeness – the way from the target population to the realized sample – and the measurement process – the way from the latent construct of interest to the actual

[1.1] This chapter is based on:
Funke, F., & Reips, U.-D. (2007). Datenerhebung im Netz: Messmethoden und Skalen [Data collection on the Web: Methods and rating scales]. In M. Welker & O. Wenzel (Eds.), *Online-Forschung 2007: Grundlagen und Fallstudien* (pp. 52–76). Köln: Halem.

response. These influences and kinds of error that may harm data quality are basically the same for paper-based studies and Web-based studies. However, due to the characteristics of the medium, there are some peculiarities about Web-based questionnaires, measurement devices, and rating scales that may influence the given answers.

1.2 Technical basics of Web-based research

This section focuses on the technical aspect of Web-based research and on the impact of technical variance on data quality, and data quantity. In addition, basic methods of Web-based measurement (i.e. data collection, data storage, and data processing) are summarized, for each of which there are various technical ways of implementation. Most ways of implementation are equivalent with regard to the possibilities of questionnaire design, even though special technologies are necessary to realize certain designs.

1.2.1 Technology and variance

In questionnaires (computerized as well as paper-based) visual design may seriously affect data quality. It is therefore crucial to keep the variance in display of the questionnaire on the participant's screen as low as possible. In contrast to studies conducted in a laboratory setting, it is almost impossible in practice to control the exact display in the respondent's Web browser. There are three main sources of involuntary differences in display on the respondent's side: hardware, software in general, and especially the Web browser.

First, it is *hardware* that determines what respondents exactly see. Screen resolution – the number of pixels a screen is made of, e.g., 1280 pixels in width by 800 pixels in height – and the actual size of the viewable area in the browser window determine if a Web page is fully displayed or if the respondent has to scroll to be able to see and evaluate the complete content and all stimuli. Also, the physical size of the screen determines the *absolute size* of the content. Given the same screen resolution, stimuli are far smaller on a 15" screen than on a 21" screen. Finally, screen and color settings – both influencing brightness and color display – can introduce additional variance. While the actual screen or window resolution can be read out automatically (using JavaScript), the size of the screen has to be asked for directly. The exact display of colors cannot be determined at all.

Second, the *software* on the respondent's computer determines if and how the content of a Web page is displayed. The correct display of certain elements (e.g., videos or sound files, special stimuli or rating scales) is only possible if the necessary software and plug-ins (programs integrated in another piece of software to extent its functionality) are installed and enabled.

The third source of variance is the *Web browser* used for participation in a study. In contrast to printed questionnaires, the exact appearance of Web questionnaires on the participant's screen cannot be determined precisely. There is always variance because of a large number of browsers (e.g., Internet Explorer, Firefox, Opera, Safari; for an overview see http://browsers.evolt.org) in different versions. Despite existing standards, some browsers interpret the HTML source code of a Web page in different ways and apply own style elements (see Reips, 2002b, 2002c; Schmidt, 2007). Dillman, Smyth, and Christian (2009) summarize studies suggesting that the resulting visual differences may have a significant effect on the given answers. Even if utmost caution is exercised and browser detection is used to eliminate known differences between browsers, there is still variation in display introduced by different Web browsers. The usage of low-tech approaches (Buchanan & Reips, 2001; Reips, 2002c) may resolve some of the compatibility problems. It is advisable, though, to consider the browser that is used for participation in a study and the available software as background variables.

1.2.2 Collecting data

The browser window is the visual interface where the respondent and the researcher virtually meet in cyberspace. The technical basis for sending, retrieving, and saving data by using a Web browser are interactive forms where all rating scales are embedded, too. The basic way to structure the content of Web pages is HTML (Hyper Text Markup Language). At present, however, the possibilities of HTML are basically restricted to static designs – the advent of HTML 5 will bring about substantial changes. More dynamic elements (e.g., immediate feedback), multimedia stimuli (like videos or audio files), demanding tasks (e.g., sorting or ranking procedures; see Neubarth, 2006), and fancy rating scales require the use of more advanced technologies. These technologies (e.g., JavaScript or plug-ins like Flash) additionally allow the unobtrusive collection of process data or paradata like response times, single keystrokes, mouse clicks or mouse movements. However, to take advantage of these possibilities, the respective technology has to be available on the respondent's computer. On the other hand,

the availability of a certain technology being a prerequisite for participation can also lead to technology-based dropout.

1.2.3 Storing data

Data entered in a Web questionnaire has to be sent from the respondent's browser and saved to make them available for further statistical analyzes (see Reips & Stieger, 2004). The most common way is to save data on a Web server, either in a *log file* or in a *database*.

Storing Data in a Log File. Log files keep records of all page requests on a Web server. These records include who (respectively which IP address) requested which information (e.g., a certain Web page or file) at which date and time with which Web browser (and operating system) from which Web page (the referrer) and if there was no error in the data transfer. Technically, the storage of data in a log file is realized with the HTTP request method GET. Here, all variable values – plus an unique user ID needed to match data afterwards if the questionnaire consists of several pages – are added to the URL of the next Web page (e.g., http://samplesurvey.net/page_2.html?variable_1=2&variable_2=5&id=1234) and saved in the log file.

One disadvantage with data storage in log files is the fact that log files cannot simply be imported in statistical software packages and that they contain many process data not needed for statistical data analyzes. So, log files have to be modified and cleared of superfluous entries (e.g., with Scientific LogAnalyzer, part of the iScience Server; http://iscience.eu) to make them available for analyzes.

Storing Data in a Database. Storing data in a database has essential advantages in comparison to the use of a log file. First of all, data do not have to be modified before analyzes and professional statistical software can directly access a database and import data. All answers are directly stored in a predefined data table. Second, data already given in a survey – in panel designs one can even use data from previous studies – can easily be used later in the questionnaire. Thus, conditional filtering, consistency checks, and personalization can be realized. Technically, the HTTP request method POST is used to save data in a database. The transferred data are not appended to the URL, and PHP or similar means of programming have to be used to process data. PHP is a server-sided technology and works regardless of the software running on the respondent's computer. Thus, no additional resources for generating individual and dynamic questionnaires are needed on the participant's

computer. A main advantage is that PHP produces plain HTML code that is less prone to error than other methods. Like all server-sided technologies PHP does not increase the risk of technology-based dropout.

There are three main disadvantages if databases are used to store information. First, there can be delays in loading times of Web pages, especially if data from many participants are to be saved simultaneously (e.g., directly after a large number of participants has been invited to participate in a study). Secondly, the use of databases requires some expertise, especially with complex study designs. Thirdly, there may be security issues with sensitive data as information in databases can be subject to attacks.

1.2.4 Processing data

In computerized questionnaires, page contents do not have to be static as is the case in printed questionnaires. Data already collected in a study can be used to modify the questionnaire. With dynamic designs, the *structure* of a study has been fixed in advance but the exact *content* that is displayed depends on previously given answers. Examples are *filtering* (skipping sections or items or branching), *randomization* of items or response options, *personalization* (e.g., using previously given information like the respondent's name), and *consistency checks*. Depending on where modifications should be implemented – on the respondent's computer or server-sided – there are different possibilities.

Processing Data on the Respondent's Computer. The participant's Web browser can be used to make questionnaires more dynamic. This can be done either by using software respectively browser plug-ins or by a combination of HTML and software like JavaScript. In this way, direct, pre-determined interactions between the respondent and the questionnaire – meaning the researcher – can be realized without having to contact the Web server. The randomization of items, the measurement of response times using JavaScript, the validation of certain types of answers (e.g., if numeric answers are within a certain interval) serve as an example. On the other hand, possibilities are quite limited in comparison to server-sided approaches.

Processing Data on the Server. The second way to make questionnaires more dynamic is the use of server-side methods, frequently implemented with PHP. Thus, complex operations can be performed and pieces of information already available in databases can be used to generate Web pages. A disadvantage of server-sided data processing is the fact that

modifications and feedback can only be realized if the respondent proceeds to the next page. So, direct feedback and modifications cannot be realized using server-side approaches.

A relatively new development is AJAX (acronym for Asynchronous JavaScript and XML), which allows retrieving additional information from a Web server without the respondent having to submit a page. Thus, rich interactions combining the advantages of server-sided and browser-based methods can be taken advantage of. For this, however, JavaScript has to be enabled in the respondent's browser.

1.2.5 Low-tech paradigm

All elementary functions needed to build Web-based questionnaires can be realized with static HTML. Many other important extensions, though, can only be realized by using server-sided solutions like PHP or Pearl. Especially user identification with login code or session IDs, conditional filtering, randomization to experimental conditions, server-sided measurement of response times, mandatory questions, plausibility checks, and even cookies can easily be implemented by using a combination of HTML, PHP, and a data base.

To take advantage of some methods – like client-sided measurement of response time, plausibility checks during data entry or some dynamic elements – advanced software is necessary. However, this software may not be available on the respondent's computer or may have been disabled due to privacy or safety concerns. This may lead to questionnaires with restricted functionality or even not working questionnaires. Technical demands, however, should not stand in the way of participation. It has to be carefully considered if the promises of advanced technologies overweigh the advantages of a robust low-tech implementation. Low technical requirements help keep dropout rates low (Schwarz & Reips, 2001) and prevent technology-caused sampling errors (Buchanan & Reips, 2001).

1.3 Reactive Web-based research and rating scales

There is a large variety of reactive methods for Web-based self-administered questionnaires. All basic question types and rating scales – open-ended questions, single choice questions, multiple-choice questions as well as simple visual analogue scales – can be implemented straightforward with HTML elements alone. Certain rating scales – like slider scales and visual analogue scales with markers – require the availability of technologies on the respon-

dent's computer. Where possible, it should be checked non-reactively in advance which technologies are available to use low-tech methods where appropriate.

Just as in paper-based questionnaires, the presentation of multiple items on a single page can lead to context effects with the response to a certain item influencing the answer to the following. Additional disadvantages are that response times cannot be determined for single items but only for the whole Web page, nor can the exact point of dropout be determined (Reips, 2002b, 2002c). Both pieces of information can be helpful to infer on item difficulty. The presentation of only one item per Web page is therefore recommended (see also Reips, 2002c for further standards of Web-based experimenting).

One important difference to printed questionnaires is that respondents are completely restricted to the options available in the questionnaire form. For instance, it is impossible with computerized rating scales to check more than one response option with closed-ended rating scales realized with radio buttons. It is not possible, either, to give a response that is between the available options or to write further clarifications as can be done on a printed questionnaire. On the surface, this leads to perfect data sets without irregularities, but in some cases respondents are deprived of the possibility to communicate important clarifications. So, problems with the construction of a questionnaire can easily be overlooked, and extensive pretesting of a questionnaire in a laboratory setting should therefore be considered a standard procedure.

The following sections focus on methodological problems that may arise in connection with Web-based questionnaires. As seen before, there is a considerable amount of variance introduced by technology. Recommendations for good practice are given wherever possible. We now take a closer look at nonresponse, which is of great relevance in Web-based research and can be turned into practical use as an indicator for problems with certain items as well as questionnaire design.

1.3.1 Nonresponse and dropout

When doing research with printed questionnaires that are to be send back by mail, respondents are likely to send only questionnaires where all or at least most items are answered. Data from respondents who did not complete the questionnaire are not collected, and every questionnaire not sent back is regarded as *unit nonresponse*. In Web-based research, data can be recorded during the process of participation. Especially with multi-page question-

naires, data are not stored on the last page of the questionnaire but continuously when the respondent proceeds to the next page. Therefore, data sets from respondents who broke off participation and responded only partially are available for analyzes, too. As a result, more (though incomplete) data sets are obtained than with printed questionnaires. Response rates, dropout rates, item nonresponse, and the point and time of dropout can be used to help understand the way questionnaires are processed and to assess the quality of a questionnaire (e.g., by identifying difficult items or technical problems). Analyzes of nonresponse (see Bosnjak, 2001) is especially important in Web-based research as it can be a major source of bias.

Two specific groups of respondents providing no meaningful answers are *lurkers* and *click-throughs*. A lurker is someone who goes through the whole questionnaire without answering a single item. In paper-based questionnaires, this would correspond to a person who sends back an unanswered questionnaire (which rarely happens). Lurkers can be individuals who want to go through a questionnaire in advance before answering the items in a second attempt or persons who have already taken part in the study and want to re-read the questionnaire. This type of lurking can only occur in studies with unrestricted access where participants are not identified (either by a personal login code, by the use of cookies, session IDs or other software-based or hardware-based identification techniques). But lurkers can also be persons who have a substantial (academic or professional) interest in the study but who are reluctant to reveal their personal views.

On the surface, lurking can be avoided by making all items mandatory. But this does not get to the bottom of the basic motivation of lurking and may lead to reactance and biased estimates (Stieger, Reips, & Voracek, 2007). Furthermore, mandatory items may produce another unwanted type of participant, *click-throughs*. Click-throughs are participants who provide meaningless answers to every (mandatory) item. Click-throughs can easily be identified in computerized data collection by controlling response times. If those times are too short, it is unlikely that ratings reflect elaborate and meaningful answers. A second way to identify this type of participant is to look for certain response patterns (like always selecting the same answer even with contradicting items).

1.3.2 Open-ended questions

The answers to open questions can be collected with text fields – presented in a single line – and with text areas – that can cover several lines on the screen – for longer texts (see Figure

1.1). Both types can be restricted to a certain number of viewable characters and to a certain number of characters that can be entered. The restriction of the number characters to a maximum is recommended in order to avoid problems with too long texts that may corrupt the log file or the database where data are stored.

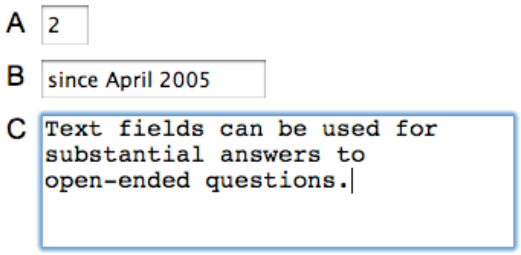


Figure 1.1. Assessing open-ended questions with single-figure text field (A), text field for multiple characters (B), and text area (C).

Aside from practical considerations – if just a numeric value, a specification or extensive information is asked for – it has to be pointed out that the size of the text field or text area exerts an influence on the reported data. Its size may even influence the measurement of quantitative variables like the monthly income (see Fuchs, 2005; Fuchs & Couper, 2001). When assessing qualitative data in self-reports, the visible size of the text area – even if there are no restrictions of the number of characters that can be entered – may have an influence on the number of words as well as on the number of topics mentioned (Christian & Dillman, 2004).

1.3.3 Single choice questions

Closed-ended dichotomous or polytomous items can be asked for by using HTML radio buttons (see Figure 1.2) or menus (see Figure 1.3). The basic problem with radio buttons is that they cannot be de-selected. Birnbaum and Reips (2005) recommend putting an extra radio button in front of every item, which also points to items that have not been answered. Providing a response option for explicit nonresponse can have a positive effect on data quality, too (Joinson, Woodley, & Reips, 2007).

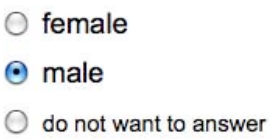


Figure 1.2. Single choice with radio buttons.

The two types of menus that can be realized with plain HTML are drop-down menus (see Figure 1.3 A) – where response options are only available after the menu has been clicked on – or scroll-down menus (see Figure 1.3 B) – where a limited number of response options is visible in a scrollable area. When using drop-down menus the first option visible should not be a valid response option but instructions on how to use the scale. Otherwise, nonresponse cannot be distinguished from the choice of the initial response option (see Reips, 2002a; Birnbaum & Reips, 2005), which may lead to a systematic overestimation of the first answer.

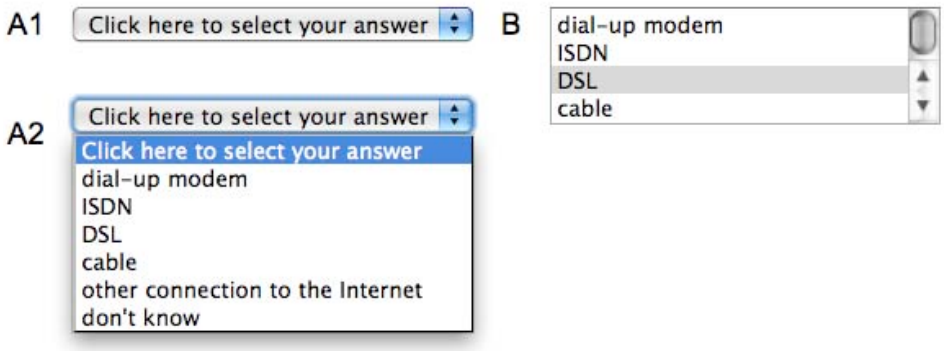


Figure 1.3. Single choice with drop-down menu, after load (A1), display of response options after menu has been clicked on (A2), and scroll-down menu (B).

When designing response scales for closed-ended questions, the order of presentation and the visibility have to be considered. Response order effects – primacy effects in the first place with visual presentation of items (e.g., Krosnick, 1991) – may occur resulting in an overrepresentation of those response options that were presented earlier. Also, if only a limited number of response options are visible – like with scroll-down menus or with extensive options in a drop-down menu – chances are that one of these options is selected at a disproportionally high rate (see Couper, Tourangeau, & Conrad, 2004).

1.3.4 Multiple-choice questions

If more than just one response option can be applicable (e.g., if fields of interest are asked for), there are two possible ways of implementation that can be realized with plain HTML. The first possibility is to use a checkbox for each response option that can be selected and de-selected (see Figure 1.4 A). If a large number of response options is presented, grouping answers can be helpful and raise the total number of selected items (Smyth, Dillman, Christian, & Stern, 2004, 2006b). This is an implicit approach where each item that is not checked is regarded as being non-applicable.

An unchecked checkbox can be ambiguous, though. It can either mean that the respondent has chosen not to select it – which would correspond to a no – or that the respondent overlooked the response option. The second possibility for multiple-choice questions is an explicit approach with both a yes and a no option offered for each item (see Figure 1.4 B), realized with HTML radio buttons. The explicit assessment of multiple-choice questions has a positive influence on the number of selected items (see Smyth, Dillman, Christian, & Stern, 2006a). Furthermore, cases of item nonresponse can be identified.

Figure 1.4. Implicit multiple-choice with checkboxes (A) and explicit multiple-choice with radio buttons.

1.3.5 Discrete rating scales

Discrete categorical rating scales offer a relatively small number of ordinal response options. These rating scales can have verbal or numerical labels (see Figure 1.5), or they can be labeled with pictures (see Jäger, 2004, for a rating scale made of equidistant smileys). But even ratings on fully anchored rating scales – where every response option is labeled – are not robust to effects of visual presentation. Horizontal or vertical presentation (see Chapter 5) or presentation in multiple columns can affect the given ratings (see Smyth et al., 2004). At any rate, the response options should always be presented at the same distance. Especially wordy labels can lead to unequal gaps, which can influence the distribution of answers (Funke & Reips, 2006).

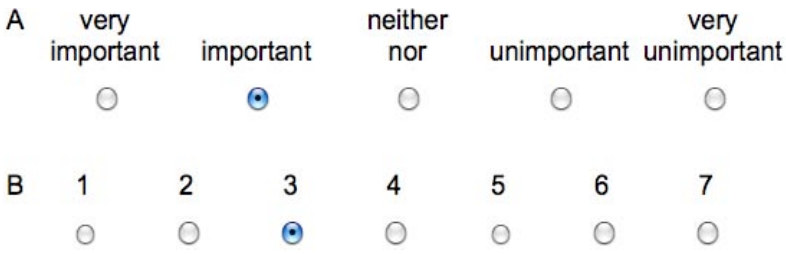


Figure 1.5. Fully labeled 5-point rating scale with bipolar verbal anchors (A) and unipolar numeric 7-point scale (B).

Matrix Rating Scales. Item batteries of discrete rating scales can also be presented in the form of a matrix (see Figure 1.6) where several items are to be rated using the same rating scale. This form of presentation is space-saving as the labels of the rating scale are only presented once. However, if very long item lists are presented, matrix questions may become confusing and discouraging (Gräf, 1999). In addition, this from of presentation may lead to undesired context effects.

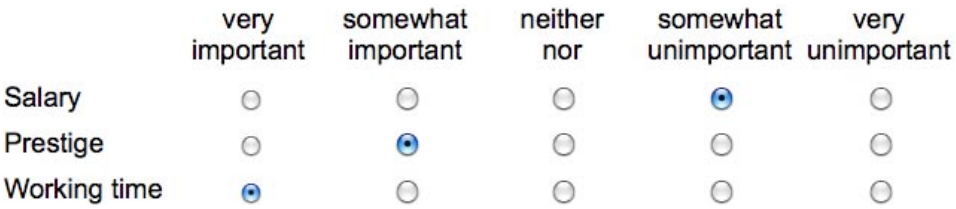


Figure 1.6. Matrix rating scale.

Semantic Differentials. If *different dimensions* of a *single construct* are to be measured, semantic differentials (see Figure 1.7) are recommended. Each dimension is evaluated on a rating scale the ends of which are anchored with opposing descriptors of the respective dimension (e.g., *exciting* and *boring* for the dimension excitement). All dimensions should be presented on the same page in order to underline that they are interconnected and respondents are expected to draw a sound picture regarding all dimensions. Any closed-ended rating-scale can be used with semantic differentials (see Chapter 4 for a comparison of semantic differentials made of 5-point scales versus visual analogue scales).

Figure 1.7. Semantic differential.

Slider Scales. Slider scales (see Figure 1.8) are graphical rating scales that cannot be realized with standard HTML and have no equivalent in printed questionnaires. Respondents indicate their response by clicking with the mouse on a movable handle, holding the mouse button, moving the handle, and releasing the mouse button when the desired position on the

Figure 1.8. Slider scale.

scale is reached. In most cases, a slider scale is no continuous measurement device, and only a limited number of discrete values – mostly labeled with markers – is offered (data from a Web experiment comparing slider scales to radio button scales are reported in Chapter 5).

Overall, it is doubtful if there are any methodological advantages to be gained from slider scales. First, there is no low-tech implementation of slider scales, which raises the risk of dropouts. Then, there is the question of where the slider should be positioned in the first place. If the handle is positioned somewhere on the scale, item nonresponse cannot be identified. If the handle is outside the scale, respondents could perceive it as an anchor stimulus representing typical cases, which might influence their ratings. In addition, the use of sliders is more demanding than that of other rating scales, which might raise respondent burden, at least for certain populations (see Funke, Reips, & Thomas, in press). On the other hand, there are only two *potential* advantages. First, the use of a second type of discrete rating scale in a long study could renew the respondent's attention. Second, with the help of slider scales a larger number of discrete values could be offered on the same screen area as compared to radio button scales that would take up much more space. However, if one wants to offer respondents a large number of response options to choose from, the use of visual analogue scales should be considered.

1.3.6 Visual analogue scales

A visual analogue scale (VAS; see Figure 1.9) is commonly made of a plain, horizontal line where only the ends have verbal, numerical or graphical anchors. Respondents indicate their answer by placing a mark at some point on the rating scale. In 1921, Hayes and Patterson described VASs in an academic context for the first time, but similar rating scales made of horizontal lines may have been used as early as the beginning of the 19th century (see McReynolds & Ludwig, 1987). With VASs, respondents do not choose between discrete categories but express their ratings on a continuous scale. VASs are especially recommended if small differences are to be detected in within- or between-subjects designs. VASs are well known in paper and pencil based research (especially in the medical sector), but there are only few studies that deal with VASs in computer-mediated research or in Web-based research (e.g., Gnambs, 2008). Such research is needed, though, as the use of VASs

Figure 1.9. Visual analogue scale.

is likely to become much more frequent in the light of the recent development of software that runs VASs on computers (Marsh-Richard, Shannon, Mathias, Venditti, & Dougherty, 2009) and on the Web (Reips & Funke, 2008; see also Chapter 2).

With computer technology, resolution of VASs can reach very high levels; in Web-based implementations of VASs, every pixel in length corresponds to a possible data point. Under technical aspects, VASs generally require more sophisticated methods like JavaScript or Flash that have to be installed and enabled on the respondent's computer. For instance, VASs generated with the free Web service VAS Generator (for a description see Reips & Funke, 2008) are based on JavaScript. If JavaScript is not available on the respondent's computer, a rating scale made of radio buttons is automatically displayed as a substitute.

VASs have the important advantage that very small differences can be detected as measurement error is reduced (see Funke & Reips, 2010a) and that data approximate the level of an interval scale (Reips & Funke, 2008). Furthermore, there is evidence that test-retest reliability is higher with VASs than with categorical scales (Funke & Reips, 2007a). Couper, Tourangeau, Conrad, and Singer (2006) observed a negative influence of VASs on completion rate and response time, but it remains unclear if this is due to the technical implementation in Java – which is far more demanding than JavaScript – or owed to VASs itself. To lower technical requirements, there is the possibility to program *simple VASs* with HTML means alone. This can be realized, for instance, if every pixel in length is made of a small graphic file. However, after a click on a simple VAS no persistent marker appears and the respondent is instantly redirected to the next questionnaire page. The limitations of this low-tech solution are that ratings cannot be altered, that the lack of visual feedback may raise task difficulty, and that only one item per page and no semantic differentials can be displayed.

In contrast to slider scales – even though there are similarities in appearance – VASs offer four main advantages: In neutral state – directly after load of the Web page – there is no marker visible that could unintentionally serve as an anchor, nonresponse can easily be identified, handling is less demanding, and data are not just ordinal but on the level of an interval scale.

Overall, VASs constitute a highly appealing rating scale for Web surveys and mode-independent measurement (see Gerich, 2007). However, some theoretical advantages still lack robust empirical foundation. More research is needed within diverse populations – espe-

cially on measurement error and test-retest reliability – as well as research where true values are known that can be used as benchmarks.

1.4 Conclusion

The general advantages of computerized data assessment (e.g., conditional presentation of items, branching, filtering, consistency checks, and randomization) can easily be used for Web-based studies. In contrast to studies conducted in a laboratory setting, Web-based research offers a number of additional advantages like access to large and geographically distant populations. All rating scales available in printed questionnaires can also be implemented in Web-based questionnaires, in most cases even by low-tech means. Furthermore, additional paradata (e.g., response times) can be made available at no costs which can be used to infer on cognitive processes, to identify respondents that are not serious and to estimate data quality. Even data from incomplete questionnaires can be used to improve questionnaires.

Technology – that has to be mastered by the research as well as by the respondent – can matter, though, and it may have a major influence on the process of data collection. Especially if the respondent's Web browser does not meet technical demands, this can result in dropout biasing estimates. Tempting new possibilities of data collection like dynamic elements or demanding new rating scales should only be used after careful and extensive pretesting as even small changes in rating scales may affect ratings. Regarding the technical implementation of a study, responsible researchers should carefully consider if the potential advantages of new methods outweigh the risk that are likely to occur if the low-tech paradigm is disregarded. Not everything that can be done should be done.

Overall, Web-based methods are a valuable contribution to already existing research methods provided that methodological knowledge is taken into account and questionnaires are carefully designed. Especially complex experimental designs with rich stimuli can easily be realized. In combination with access to large and diverse populations, Web-based research opens up a multitude of new possibilities for psychological research.

2

Interval-Level Measurement With Visual Analogue Scales in Internet-Based Research: VAS Generator

2 Interval-Level Measurement with Visual Analogue Scales in Internet-Based Research: VAS Generator[21]

The present article describes VAS Generator (www.vasgenerator.net), a free Web service for creating a wide range of visual analogue scales that can be used as measurement devices in Web surveys and Web experimentation, as well as for local computerized assessment. A step-by-step example for creating and implementing a visual analogue scale with visual feedback is given. VAS Generator and the scales it generates work independently of platforms and use the underlying languages HTML and JavaScript. Results from a validation study with 355 participants are reported and show that the scales generated with VAS Generator approximate an interval-scale level. In light of previous research on visual analogue versus categorical (e.g., radio button) scales in Internet- based research, we conclude that categorical scales only reach ordinal-scale level, and thus visual analogue scales are to be preferred whenever possible.

2.1 Introduction

Internet-based research has become a new procedural option in psychology and related sciences, and many researchers are looking for tools, methodologies, and services that support this type of research. If such tools for Internet-based research are offered for use via the Web (not via download and install), they are called *Web services*. Such Web services are available wherever there is an Internet connection and a Web browser.

Web services for Internet-based research include *collections of Web studies* for recruitment and archiving (e.g., the "Psychological Research on the Net" list by John Krantz, and the "Web Experiment List" at genpsylab-wexlist .unizh.ch/; Reips & Lengler, 2005), tools for *Web surveying* (e.g., surveymonkey.com or surveyWiz; see Birnbaum, 2000b), *Web experiment generators* (e.g., WEXTOR at wextor.org; see Reips & Neuhaus, 2002), *Internet-based tests* for inclusion in studies (e.g., b5 at iscience.eu), *Web log analysis tools* (e.g., Scientific LogAnalyzer; Reips & Stieger, 2004), and *portals* that link to all related services (e.g., the iScience server at iscience.eu).

[2.1] This chapter is based on a presentation given at the 36th Annual Meeting of the Society for Computers in Psychology, Houston, TX, on November 16, 2006. It has been published as:
Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods, 40,* 699–704.

The present article describes a first example of a new set of tools that can be termed *Web measurement device generators*. VAS Generator (www.vasgenerator.net) is a free Web service for creating a wide range of visual analogue scales that can be used as measurement devices in Web surveying and Web experimentation, and also for local computerized assessment. VAS Generator and the scales it generates are platform-independent, meaning that the scales can be created from and used on all types and versions of operating systems with Web browsers that support the underlying universal languages of HTML and JavaScript. The resulting scales can easily be added to surveys and experiments generated with other Web services, such as surveyWiz (Birnbaum, 2000b) and WEXTOR (Reips & Neuhaus, 2002).

2.2 Visual analogue scales

Visual analogue (or analog) scales (VASs) are continuous measurement devices (see, e.g., Flynn, van Schaik, & van Wersch, 2004). In 1921, VASs were described for the first time (Hayes & Patterson, 1921). However, this type of scale was not seriously examined before 1969 (Aitken, 1969). In some respects, Aitken's basic findings are still state of the art, since – unlike on other measuring instruments – little research has been conducted on this type of scale (for an overview, see Flynn et al., 2004). Practical concerns are the major reason for this lack of research. In paper-based VASs, a lot of time and effort are required for reading the data: The exact position of each marking has to be determined by hand. Of course, the situation changed considerably with computerization. With the rise of Internet-based research, VASs have become a measurement device that could be used widely without practical drawbacks (Couper, Tourangeau, Conrad, & Singer, 2006; Funke & Reips, 2006). The wider use of VASs may even solve some issues related to data quality with other online measurement devices (Funke & Reips, 2006; Reips, 2002a; Smyth, Dillman, Christian, & Stern, 2006).

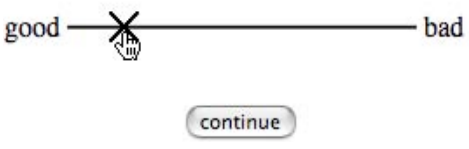


Figure 2.1. A visual analogue scale generated with VAS Generator.

A VAS consists of a line and two anchors, one at each end. The anchors often consist of verbal materials that mark opposite ends of a semantic dimension (e.g., *good* and *bad*).

However, the anchors may also be pictures, or even sound files. Visual anchors, such as smileys, are also often used with participants who may not fully grasp the meaning of verbal materials—for example, with preschool children. See Figure 2.1 for an example VAS generated with VAS Generator.

In general, VASs are considered a reliable instrument for valid measurements. However, recent research has mostly been limited to comparisons between VASs and categorical scales (Likert-type questions) in a paper-andpencil environment. The majority of studies have been conducted in the medical sector, in which reliable detection of small changes in status is appreciated a lot, especially in the fields of pain and fatigue research (e.g., Bellamy, Campbell, & Syrotuik, 1999; Brunier & Graydon, 1996). A notable exception, conducted by Gerich (2007), compared VASs and 5-point categorical self-control scales not only in a paper-and-pencil format, but also with computerassisted self-administered interviews.

VASs provide researchers with a number of advantages. In comparison with discrete scales, measurement by a VAS is more exact, and the scale needs less explanation for the research participants (e.g., smiley-face scales in studies with children). In a previous study, we were able to show that online radio button scales do not show linear correspondence to VASs (Funke & Reips, 2006). In that study, two experiments were conducted, comparing categorical scales of 4, 5, 7, 8, and 9 points with a VAS of medium length (200 pixels). The study presented here further investigated whether the VAS format diverges from the interval level and whether the length of a VAS may be a boundary condition for such an effect. First, we describe VAS Generator.

2.3 Creating and implementing a scale with VAS Generator

Just five steps are required to generate an individual VAS for use on a Web page. In a simple HTML form available at www.vasgenerator.net (see Figure 2.2), the researcher modifies the essential graphical parameters (length, color, width, and type of marking) and the verbal anchors. Below we explain the five necessary steps.

***Step 1: Defining Core Scale Parameters*.** Several mandatory parameters are set at default levels and can be adjusted according to one's needs. These parameters are the length, anchors, color, and marker. The length of the VAS by default is 200 pixels, a length that will be displayed fully without problems in most Web browsers and with most currently used monitor resolutions. The verbal anchors located at each end of the VAS have a preset maxi-

mum length of 50 characters, but this value can be adjusted when pasting the code to a Web page. The scale's color can be changed from black (which is the default value) to white, and the width from medium (3 pixels) to light (1 pixel) or boldface (5 pixels). The marker that will appear after clicking on the scale, representing the user's value, can be modified from a cross to an arrow, point, or line. Measurement resolution—that is, the number of (discrete) points to click—can also be adjusted.

Figure 2.2. VAS Generator's main window.

After confirming these parameters by clicking on the "generate/modify VAS" button, a working preview of the scale is displayed in paragraph 2 on the same Web page.

Step 2: Testing the Scale Draft. The VAS that has been generated can be clicked on in order to display the value (i.e., the position of the marking, measured in pixels from the left-hand side) in a pop-up window.

Figure 2.3 shows a bold, black VAS that is 400 pixels in length. The left end is anchored with *strongly agree* and the right with *strongly disagree*. A cross acts as the indicator of the respondent's value (here, at 115 pixels).

Figure 2.3. Read out of rating value in feedback window.

Step 3: Downloading the Basic Scale Materials. If the VAS satisfies one's needs, all of the basic files required to include the VAS on a Web page (i.e., the JavaScript code and picture files, with an additional instruction for offline use) can be downloaded as a compressed file in .zip format. Links to free software that can be used for decompressing ("unzipping") the file are provided. All files in the decompressed folder need to be stored in the same folder as the Web page that will contain the newly generated VAS.

Step 4: Downloading the Customized Scale. The VAS that was customized in the first two steps can be displayed in a separate browser window by clicking on the "go to VAS and download" button. The source code for this window is downloaded by simply saving the page from the browser menu into the folder with the basic files (Step 3) and changing the file extension from .php to .html. The basic file folder should now contain an HTML document, two JavaScript files (ending in ".js"), and an "images" subfolder containing the picture files needed for displaying the scale. The JavaScript code for the rating scales was originally developed by Walter Zorn and Timo Gnambs and is distributed as open source (timo.gnambs.at/downloads/zip/ratingscales.zip).

Hence, there are two options: The VAS can either be pasted easily into an existing Web page or serve as the basis for a new survey. Pasting the VAS into an existing Web page will likely be the more frequent way of proceeding. Therefore, the related editing procedure in HTML is described in Step 5 below.

Step 5: Implementing the Scale on a Web Page (Includes the Resulting Code). To integrate the VAS with an existing project, the source code of the downloaded page needs to be modified. Only three parameters (printed in capitals in Figure 2.4) have to be adjusted: The name of the current page ("YOUR_PAGE_TITLE") that is displayed at the top of the browser window, the name of the following page ("YOUR_NEXT_PAGE.html") and – this is the most important step if one uses several VASs in one survey – the name of the current scale ("THIS_VAS_NAME").

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 TRANSITIONAL//EN">
<html>
    <head>
        <title>YOUR_PAGE_TITLE</title>
    </head>
    <body>
<!-- "wz_dragdrop.js" and "ratingscale.js" have to be included immediatly after "<body onload='next_page()'>" -->
        <script type="text/javascript" src="js/wz_dragdrop.js" title="Dragdrop Bibliothek (c) Walter Zorn"></script>
        <script type="text/javascript" src="js/ratingscales.js" title="Ratingscales Addon (c) Timo Gnambs"></script>
        <p>
            <table align="center" border="0">
                <tr>
                    <td align="center">
                        <img src="pics/pix.gif">
                        <form name="vas_data" action="YOUR_NEXT_PAGE.html" method="get">
                            <script type="text/javascript">
scales.scale('THIS_VAS_NAME', 'click', 400, ['pics/line20.gif', , 10], ['pics/00.gif'], ['strongly agree', 'strongly disagree']);
                            </script>
                    </td>
                                    </tr>
                <tr>
                    <td align="center">
                        <input type="submit" value="continue" />
                        </form>
                    </td>
                </tr>
            </table>
        </p>
<!-- "<script type="text/javascript">scales.init('value');</script>" has to be immediately before "</body> -->
        <script type="text/javascript">scales.init('value');</script>
    </body>
</html>
```

Figure 2.4. Placement of customized code for one's particular VAS in an HTML file.

2.4 Analyzing data from a VAS

Data from the VAS contain the following information: a “name” that was given to this particular item (in order to separate entries from those related to other items) and a “value” that reflects a participant’s rating on the scale. The value is read out automatically, with an accuracy of one pixel. This means that the greater the length of a VAS, the more diversified the range of possible values. In other words, the precision of the rating is only as high as the length of the VAS permits.

The type of data analysis depends on the method used to collect data on the Web page the VAS was embedded in. For example, data would be ready for direct download if WEXTOR’s hosting feature were used. Data could also be written to a database or retrieved from the server log file, which can be analyzed with Scientific LogAnalyzer (Reips & Stieger, 2004).

2.5 Empirical test of interval level measurement

To examine whether VASs created with VAS Generator produce data on the level of an interval scale in Web studies, we conducted a Web experiment. Furthermore, we were interested in putting the scales’ robustness to the test by introducing extremely short or long VASs and by varying the mathematical format (percentages vs. ratios) of the target values.

2.6 Method

2.6.1 Procedure

In a Web experiment, participants were randomized to one of six conditions in a 2 (mathematical format) × 3 (length of VAS) experimental design. The Web experiment consisted of 30 pages per condition. The first page contained an introduction to the Web experiment and a seriousness check (Reips, 2002b). On the second page, participants were asked their gender, online experience (since when and frequency), speed of connection, and screen size. They were instructed to repeatedly identify 13 different values (percentages – e.g., 20%, 50% – or ratios – e.g., 1/5, 1/2 – depending on the mathematical format condition) in one of three length conditions: a VAS length of 50, 200, or 800 pixels. The 13 values ranged from 5% to 95% and were displayed in two reverse orders, AB and BA (see the A and B rows

in Table 2.1), always with 50% as the first item. We tried to keep a large difference between consecutive values. The underlying measurement on the VAS ranged from 0 to 100. Before and after the second block, we asked participants about three usability considerations: accuracy, precision, and interest.

Table 2.1

Sample Length Values Displayed to Participants

	Percentage Condition												
A	50	75	10	33	80	95	25	67	40	5	60	90	20
B	50	20	90	60	5	40	67	25	95	80	33	10	75
	Ratio Condition												
A	1/2	3/4	1/10	1/3	4/5	19/20	1/4	2/3	2/5	1/20	3/5	9/10	1/5
B	1/2	1/5	9/10	3/5	1/20	2/5	2/3	1/4	19/20	4/5	1/3	1/10	3/4

2.6.2 Participants

The participants were recruited from a student-based online panel at the University of Kassel that is maintained by the second author. Within 5 weeks, 439 persons took part in our study. Of the 405 participants (92.3%) who indicated serious participation, 205 were randomly sent to the percentage condition and 200 to the ratio condition. In the percentage condition, 17 participants (8.3%) dropped out during the study, and 31 (15.5%) dropped out from the ratio condition. Furthermore, 2 participants in the ratio condition did not provide even a single rating for 1 of the 13 values. Their data were excluded from analysis, so that 355 data sets (188 in the percentage condition, 167 in the ratio condition) remained for further analyses.

2.7 Results

For each individual, we computed the average ratings for the same values from the A and B orders.

2.7.1 Outliers

A standard outlier analysis revealed that overall ratings averages by 2 participants in the percentage condition and 8 in the ratio condition differed from the pack by more than three in-

terquartile ranges (extreme outliers). Seven of these 10 cases were from the 50-pixel, 2 from the 200-pixel, and 1 from the 800-pixel length condition. All extreme outliers were excluded from further analyses.

2.7.2 Interval-level measurement

For each individual, we computed the average ratings for the same values from the two orders A and B. Overall, the target values matched actual values well (see Figure 2.5). With two exceptions at the low end and two in the center, the VASs slightly underestimated the true values. This tendency was a bit stronger for the shortest VAS. Overall, the mean deviation from the target value ranged from 3.94 points for the short VAS to 2.98 for the long VAS, with 2.77 for the medium VAS (Table 2.2).

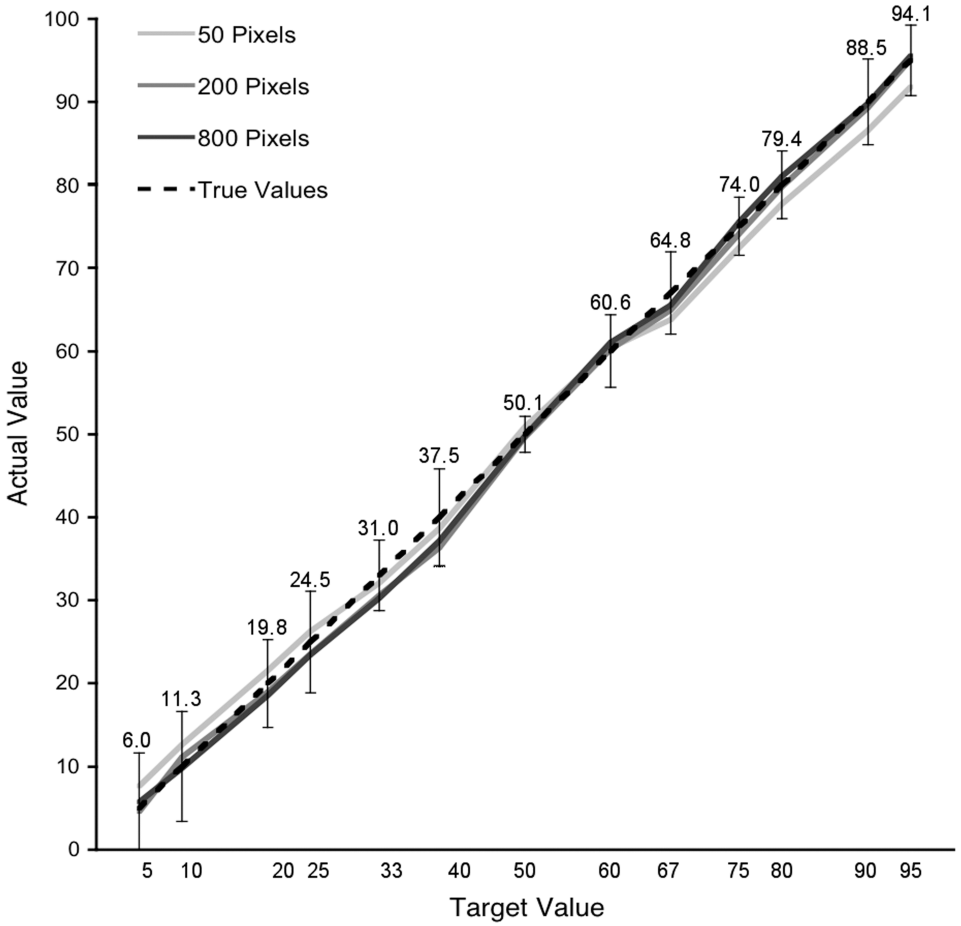


Figure 2.5. Estimated versus true values for VAS Generator VASs of various lengths. All ratings have been converted to percentages for the purposes of this figure.

Interval-level measurement implies even differences between values with even distances. Therefore, we examined whether participants produced the same differences between their ratings on each portion of the scale. The mean differences for 5-percentage-point intervals (with *SD*s in parentheses) were 4.95 (8.30) for the 5%–10% interval, 4.40 (5.54) for the 20%–25% interval, 5.26 (5.07) for the 75%–80% interval, and 5.15 (7.48) for the 90%–95% interval. The mean differences for 10-percentage- point intervals were 8.47 (7.42) for the 10%–20% interval, 12.52 (6.47) for the 40%–50% interval, 10.14 (5.24) for the 50%–60% interval, and 8.92 (6.36) for the 80%–90% interval.

Table 2.2

Mean Differences Between Target Values and Actual Values by VAS Length and Condition

VAS length	*M*	*N*	*SD*
	Combined		
50 pixels	3.94	125	2.24
200 pixels	2.77	105	1.26
800 pixels	2.98	115	1.91
Total	3.26	345	1.94
	Percentage		
50 pixels	4.03	64	2.59
200 pixels	2.75	59	1.09
800 pixels	2.73	63	1.18
Total	3.18	186	1.87
	Ratio		
50 pixels	3.85	61	1.82
200 pixels	2.79	46	1.47
800 pixels	3.28	52	2.50
Total	3.36	159	2.02

2.7.3 Analysis of potential biases when using extreme scales

We computed the differences of the average ratings from the true values. With the overall averages of these deviations from the true VAS values, we conducted an ANOVA with the factors mathematical format and length of VAS. The difference between the two format conditions was small (mean deviations: 3.18, SD = 1.87, for percentages, and 3.36, SD = 2.02, for ratios) and not significant, $F(1, 335) < 1$, n.s. Length of VAS turned out to have a highly significant effect, $F(2, 335) = 17.76$, $p < .001$, $\eta^2 = .09$, but the interaction was not statistically

significant, $F(2, 335) < 1$, n.s. Table 2.2 shows average differences between the actual value and the average VAS rating by condition.

Planned comparisons were made between the different levels of the length factor. Both main effect contrasts between the 50-pixel length condition and the other conditions were statistically significant, $F(1, 338) = 30.13$, $p < .001$ (200 pixels), and $F(1, 338) = 22.21$, $p < .001$ (800 pixels). No significant difference was found for the 200-pixel versus 800-pixel comparison, $F(1, 338) = 0.074$, n.s.

Technically, if a participant's value is off by 1 pixel on the 50-pixel VAS, an error of two units will be recorded, whereas a 1-pixel error is only 0.125 units on the 800-pixel VAS. Therefore, the larger error in the 50-pixel VAS comes as no surprise and can be explained mostly by this lack of technical precision. Furthermore, the short VAS may have fostered a tendency to repeatedly click in the middle of the scale – a behavior that would explain the shape of the 50-pixel curve in Figure 2.5.

As an aside, the large difference in dropouts between the mathematical format conditions (odds ratio = 1.87) and a four-times-higher rate of extreme outliers indicates that the ratio condition is more difficult than the percentage condition. It can be assumed that the voluntary Web mode in most Web experiments is a comparatively good indicator of participants' motivation to continue with the study (see, e.g., Reips, 2002b, 2002c, 2007). Accordingly, it seems wise to avoid the ratio format with VASs, because users seem to have less motivation when using this format, and consequently either may not answer (i.e., drop out when possible) or may produce lower-quality data if they find the situation less easy to leave and feel forced to answer, as may often be the case in laboratory studies with an experimenter present.

2.8 Conclusion

Because equal numerical intervals corresponded to roughly equal segments on the VASs and only minor aberrations from true values were found across the scale, even for extremely short and extremely long VASs, there is strong evidence that data collected with VASs are equidistant and are on the level of an interval scale. Therefore, a wide range of statistical procedures can be applied safely when analyzing data measured with VASs that were created with VAS Generator. In contrast, because there is a systematic difference between

equally spaced radio buttons and VASs (Funke & Reips, 2006), we conclude that measurements with radio button scales differ from interval level.

Although the effects were small, we found differences for the different VAS lengths on ratings of some values. However, we used an extremely long (800-pixel) and an extremely short (50-pixel) VAS that may have looked highly unusual to most participants. If on a given screen or printed page, a 50-pixel VAS is about as wide as the word “participant” on this page, then an 800-pixel VAS is about as wide as the text on this full page would be high. Given these sizes, it is surprising how well the participants performed using the 800-pixel scale. This finding implies that measurement with VASs is robust to differences in size due to different screen sizes and resolutions. Nevertheless, care should be taken in determining a VAS that is well-suited for all displays used in a study, which most often will be a medium-sized one (150–400 pixels).

Finally, as compared with hand-coding HTML and JavaScript, VAS Generator greatly reduces the effort required in generating VASs for Web-based studies. It provides all of the files needed for the use of VASs and offers preview and pretest functions. Last but not least, VASs built by means of the VAS Generator can easily be included in existing projects (e.g., in Web experiments created with WEXTOR).

3

Making Small Effects Observable: Reducing Error by Using Visual Analogue Scales

3 Making Small Effects Observable: Reducing Error by Using Visual Analogue Scales[3.1]

This paper focuses on the theoretical and empirical extent of formatting error when measuring continuous variables. In 3 independent studies we confirmed the superior properties of visual analogue scales (VASs) and show how the common way of using ordinal scales to measure continuous variables increases error. Data assessed with VASs are affected by significantly less error leading to more narrow confidence intervals and more power of statistical tests. This facilitates detecting small effect sizes that are unobservable with ordinal scales. Studies with VASs are accordingly more efficient as effects can be detected with smaller samples than needed with ordinal scales.

3.1 Introduction

Error is an omnipresent opponent for those in charge of designing high-quality studies. It seriously harms data quality, it makes analyses more difficult and it may even lead to wrong inference. Identifying and eliminating sources of error is a key interest of survey methodologists. This paper focuses on how visual analogue (or analog) scales (VASs) minimize formatting error– i.e. the difference between true value and reported value – in self-administered questionnaires. Even though there are strong arguments – like high sensitivity and continuous data – for the use of VASs, these scales are seldom used in practice. To investigate further beneficial effects of VASs in the process of data collection, we report a systematic simulation and two empirical studies. In Study 1 we use differently distributed simulated data sets to analyze formatting error arising with VASs and ordinal scales with various numbers of response options. In Study 2 we conducted a Web experiment to find out if the hypothesized formatting error with VASs is indeed lower than with 5-point scales. Finally, in Study 3, we investigate the predicted relation between the number of response options and formatting error in an open-access Web experiment where 5-point, 7-point, and 9-point scales are compared to VASs.

[3.1]This chapter is based on presentations given at the third conference of the European Survey Research Association, Warsaw, Poland, June 29–July 3, 2009, and at the 39th annual conference of the Society for Computers in Psychology (SCiP), Boston, November 19, 2009. You may refer to this chapter as:
Funke, F., & Reips, U.-D. (2010). *Making small effects observable: Reducing error by using visual analogue scales.* Manuscript submitted for publication.

3.1.1 About visual analogue scales

VASs are very simple yet powerful measurement devices for self-administered questionnaires. They are graphical rating scales consisting of a (mostly plain) line with anchors only at the very ends (see Figure 3.1). Stimuli used as anchors can be of different modality (e.g. visual or aural), they can be abstract or concrete (e.g. pictures, verbal descriptors, colors). Respondents indicate the extent of the variable measured (e.g. the consent to an item on an agree-disagree scale) by placing a mark on the line where answers are not restricted to a number of discrete response options. VASs are continuous measurement devices allowing the detection of small differences between subjects and within subjects. In computerized assessment – where dimensions of objects are given in pixels rather than in absolute units like centimeters – every clickable pixel on the line serves as a possible response option.

Figure 3.1. VAS, 250 pixels in length for computerized data collection.

Disadvantages and Solutions. One disadvantage of data collection with VASs is that the exact read-out of the given rating in paper-based studies is quite burdensome. The determination of the exact position of the marker requires a lot of time-consuming and error-prone manual processing and makes this type of rating scale especially inconvenient for large studies. This explains why these scales are mostly used in the medical sector where the importance of sensitive measurement devices outweighs the efforts of read out and where samples are usually relatively small. A large body of literature illustrates this (Bijur, Silver, & Gallagher, 2001; Myles, Troedel, Boquest, & Reeves, 1999; Seymour, Simpson, Charlton, & Phillips, 1985). In computerized data collection the automatic read-out of data is fast and accurate. By digitizing VASs these disadvantages become obsolete and the virtue of VASs can be taken advantage of for other branches of science and for large-scale surveys.

One drawback is that VASs in Web-based studies cannot be realized in plain HTML that can be interpreted by every Web browser. If certain technologies (like JavaScript, Java or Flash) are not available in the respondent's Web browser surveying with VASs is impossible. This kind of dropout is not considered to be neutral, as technological variables and personality traits can be confounded (Buchanan & Reips, 2001; Reips, 2007; Schmidt, 2007). There is

the risk of seriously biasing results if technology is made a prerequisite for participation in a study.

One convenient way of creating VASs for computerized data collection in the laboratory or in the Web is the free Web service VAS Generator (http://vasgenerator.net) maintained by the authors (see Reips & Funke, 2008). The underlying technology required in the respondent's browser is JavaScript. If JavaScript is not available, the code automatically displays equally spaced radio buttons to avoid dropout because of technology. The VAS on a Web page appears as a plain line between two anchors. To make a rating the cursor is moved on the line. The mouse (or whatever input device is used) is clicked on and a marker (here: a cross) instantly appears. Ratings can be adjusted as often as wanted by re-clicking the scale but the marker cannot be slid to a certain position. The final judgment is submitted when the button leading to the next page is clicked on.

One could argue that VASs can be substituted by using open text questions where values between 0 and 100 should be entered to express the extent of the latent variable. But this kind of assessment leads to the well-known effect of heaping or bunching (e.g. Jensen, Turner, & Romano, 1994; Tourangeau, Rips, & Rasinski, 2000) where an unreasonably high number of reports are multiples of five or ten. In none of our studies with VASs (e.g. Funke & Reips, 2006; Reips & Funke, 2008) we observed heaping or bunching. One explanation is that VASs produce a clear picture of the concept of a continuous variable in the respondent's mind. In contrast, the task of providing a number could be too burdensome so that cognitive shortcuts are used.

Measurement With VASs. In numerous studies it could be shown that mean ratings obtained with VASs do not differ from mean ratings with categorical scales, even in Web-based research (Couper et al., 2006; Funke & Reips, 2010). Averbuch and Katzper (2004) provided patients suffering from osteoarthritis pain (N = 98) with 5-point scales (with all options verbally labeled) and VASs (with only the ends verbally labeled) for rating acute pain. For each of five measurements they found very high correlations, $r^2 > .995$. However, the authors found some inconsistencies in using the scales and a considerably large range of VASs values associated with each response option. The relation between VASs and ordinal scales is linear for only about half of the participants. This finding underlines the difficulty of transforming data from VAS to categories on the individual level. Similar results have been reported for computerized studies (Funke & Reips, 2006; Gerich, 2007). Myles et al. (1999) and Myles and Urquhart (2005) found that VASs produce interval data. Hofmans and Theuns (2008) summarize that "VASs can be considered as linear scales and that the type of end anchors

used has no effect on the linearity of the VAS data" (p. 401). In a Web experiment Reips and Funke (2008) found that data from VASs approximate the level of an interval scale, even for very short and very long scales, and conclude that these scales are robust to differences in display that are common in Web-based research. Tiplady et al. (1998) examined the influence of age on the use of VASs (younger participants: mean age = 30.2 years; older participants: mean age = 70.8 years). They found that "the use of visual analogue scales is valid in both older and younger healthy subjects" (p. 65).

Despite the fact that VASs lend themselves for computerized assessment (Kreindler, Levitt, Woolridge, & Lumsden, 2003), only few studies using VAS have been conducted on the Web (e.g. Couper, Tourangeau, Conrad, & Singer, 2006; Funke & Reips, under review; Reips & Funke, 2008; Turpin et al., 2003; van Schaik & Ling, 2007). Couper et al. (2006) found more dropout and more item nonresponse as well as longer completion times with VASs realized with Java. Another study (Funke & Reips, 2010) suggests that these negative effects are caused by the Java technology and not by the scale itself and can be avoided by using more lightweight technologies like JavaScript.

3.1.2 Measurement error and formatting error

Survey methodology aims at detecting sources of error and reducing its overall effect. To specify the kind of error we tackle in this paper we combine two approaches: The concept of the *total survey error* (Groves et al., 2004) and the *question-answer process* (Schwarz & Oyserman, 2001).

Groves and colleagues (2004) describe a well-established road map in survey research that draws attention to major sources of error. The concept of *total survey error* takes many possible sources of error into account that cumulate within one survey. On the most general level one distinguishes between error of non-observation and error of observation that both affect survey statistics. *Error of non-observation* is about the representativeness of a study. It can occur at each step on the way from the target population over the sampling frame (here coverage error may occur), the obtained sample (source for sampling error), the actual respondents (nonresponse error), and post-survey adjustments (adjustment error) to the final survey statistics.

The *error of observation* is about the measurement process. It is the way from the abstract construct that should be measured over the more concrete measurement (e.g. observations

or questions in a survey) where validity can be affected, over the actual given response (where measurement error can occur), and the edited response (e.g. after post-survey adjustments like removing of outliers; source for adjustment error), to the final survey statistic. The kind of error this paper focuses on is measurement error, "the observational gap between the ideal measurement and the response obtained" (Groves et al., 2004: 51). Measurement error again can have different sources like mode of presentation (e.g. primacy effects with visual presentation and recency effects with aural presentation; Krosnick & Alwin, 1987), presence of an interviewer (e.g. social desirability even with virtual interviewers; Fuchs & Funke, 2007), questionnaire design (e.g. paging versus scrolling approach, see Dillman, Smyth, & Christian, 2009), and the source of errors can lie within respondents.

The *question-answer process* as described by Schwarz and Oyserman (2001) can be used to explain sources of measurement error that are located within respondents. The authors name five implicit demands where error can occur. Participants have to (1) understand the question, (2) recall relevant information, (3) make inferences and estimate, (4) map answers onto the available response scale, and (5) edit answers according to social desirability (see Schwarz & Oyserman, 2001, p. 129). Examples are if respondents misunderstand a question (e.g. when items are presented in a certain context; error on step 1), events are hard to remember, very rare or very frequent (e.g. telescoping effects; step 2 and 3) or if respondents are reluctant to give honest answers (e.g. answering in direction of the point of social desirability with sensitive items; step 5). The focus of the research presented is error occurring in the step 4, when problems with mapping answers onto the available rating scale may occur. It arises when respondents do not find an option on a rating scale that perfectly reflects their true value and they have to settle on a sub-optimal option.

3.1.3 Operationalization of formatting error and impact on data quality

The general description of the measurement process according to classical item response theory (e.g. Groves et al., 2004) is $Y_i = T_i + e$ where Y_i is the observed variable for respondent *i* that consists of respondent *i*'s true value T_i and an error term *e*. In the following analyses we will look at formatting error only.

For a single measurement, *e* can be either be zero or greater or smaller than zero. In the first case the answer option selected perfectly fits the respondent's true value and the reported value equals the true value. In the second case *e* is greater or smaller than zero leading to a systematic over- or underestimation of the true value resulting in *biased* estimates.

When dealing with more than just one respondent (e.g. in a sample survey) or with the measurement of multiple items (e.g. when applying Likert's method of summed ratings) the mean population value *Y* is computed as follows:

$$\sum_{i=1}^{n} Y = \frac{(T_1 + e_1) + (T_2 + e_2) + \ldots + (T_n + e_n)}{n} \quad (1)$$

Again, there can be a perfect measurement for all respondents or items leading to *e* of zero and again bias means drawing systematically wrong inference and arriving at a distorted picture of the population. The additional third case is that over- and underestimations may not be systematic but random. Thus they would even out, resulting in *e* having an expected mean of zero and a variance greater zero. This kind of error cannot be observed on the first sight, and valid inference on parameters is drawn. It has some serious but subtle impact on data quality. An increase of error leads to wider confidence intervals of survey estimates. This leads to reduced power of statistical tests and to decreased reliability. The first consequence is that larger sample sizes are needed to statistically significantly detect existing differences. Secondly, correlation coefficients are reduced by increasing error. Finally, no matter how large the sample size, small effects cannot be observed. Obviously, there is every reason to maximize efforts to minimize error.

3.2 Study 1: Simulation of formatting error with ordinal scales and VASs

The most frequent rating scales in self-administered surveys are ordinal scales. Whenever ordinal rating scales are used to measure continuous variables there are unavoidably many instances where there is an observational gap between the extent of the variable of interest and the reported value. Aim of the first study is to gain deeper understanding for how formatting error is affected by the number of response categories, if continuous variables are measured with ordinal scales. Starting point is the definition of formatting error stated above. In the first section we describe the expected extent of formatting error for a single measurement with ordinal scales and with a VAS. In the second section we use simulated data sets to estimate formatting error occurring within certain populations with different distributions of the dependent latent variable. Finally, we summarize the theoretically expected formatting error in ordinal scales and VASs as foundations for the empirical Studies 2 and 3.

3.2.1 Formatting error for single measurement

To operationalize formatting error we put it in numbers by thinking the latent variable to be measured as being *unipolar* and *continuous* running from 0 to 100.[3.2] At the left end of this continuum 0 represents the absolute absence (e.g. indicated by a verbal label like “do not at all agree” when items are to be rated or “never” when asking for frequencies) and 100 at the right end of the continuum is the absolute presence (e.g. “strongly agree” or “always”) of the variable being measured. According to this concept a measurement error of one unit means a deviation of about one percentage point from the true value. Furthermore, we assume that the true value can take any whole number between 0 and 100 and it is represented with the same precision within the respondent. The respondent can access this value with any precision and map it on the rating scale.

Formatting error is the difference between true value and the corresponding optimal value communicated on the measurement device (i.e. the scale). To compute this difference we have to determine the respective category value for each option on an n-point ordinal scale in advance. The goal is to assign a value to each category, resulting in numerically equidistant categories. We computed each category’s value, given that the left most category represents the value 0 and the right most category represents the value 100. The distance d between each category equals the total number of categories c minus one: $d = 100 / (c - 1)$. For a 5-point scale this means that the distance between categories is equal to $d = 100 / (5 - 1) = 100 / 4 = 25$. This leads to category values of 0 for the left most category, 25 for the second category, 50 for the middle category, 75 for the fourth category, and 100 for the right most category.

For illustration of how formatting error is computed for different ordinal scales and a VAS, consider the following example. Let a respondent i’s true value T_i be of a middle intensity, precisely scoring at 42. Figure 3.2 illustrates the extent of formatting error occurring with a 5-point scale, a 9-point scale and a VAS.

In the example 5-point scales overestimate the latent variable, leading to a formatting error of eight percentage points. 9-point scales slightly underestimate the true value by four percentage points. Not surprisingly, the formatting error with VAS equals zero, as there is a perfectly matching response value on the VAS that equals the true value.

[3.2] In fact this procedure leads to 101 categories. However, it also leads to exactly 100 increments. Thus, we arrive at the same distances between crucial values, e.g. 0–25, 25–50, 50–75, and 75–100 are equally large, and 50 truly marks the middle of the scale. In any case, differences are neglible and were taken into account during computations.

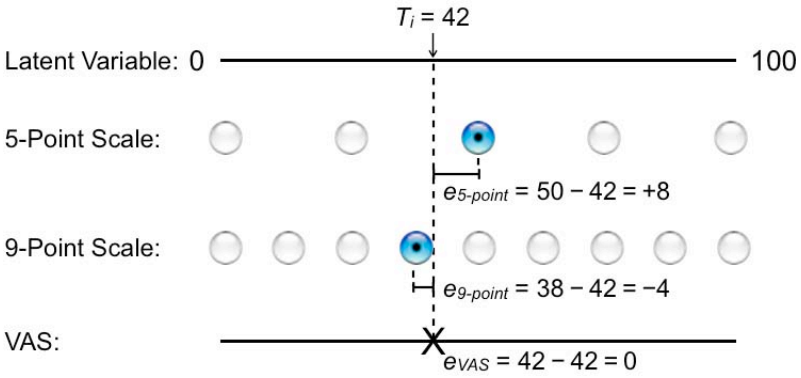


Figure 3.2. Illustration of within-category formatting error (*e*) operationalized as difference between true value (T_i) and the nearest answer option with 5-point scale, 9-point scale and VAS.

To determine the mean expected formatting error resulting for different ordinal scales we determine the formatting error for each possible true value between 0 and 100 by computing the distance between the actual true value and the closest category. Because all response options are equidistant, the mean expected formatting error approximates one fourth of the category width for every category. As there are 101 possible values (0 and 100 are included in the interval) the expected formatting error *e* equals one fourth of the category distance *d* corrected by 100/101: *e* = *d* / 4 * 100 / 101.

Table 3.1

Distance Between Categories, Maximum Formatting Error, and Expected Formatting Error with Ordinal Scales Consisting of Three to Twelve Categories

	Number of categories									
	3	4	5	6	7	8	9	10	11	12
Category distance *d*	50.0	33.3	25.0	20.0	16.7	14.3	12.5	11.1	10.0	9.1
Maximum FE	25.0	16.7	12.5	10.0	8.3	7.1	6.5	5.6	5.0	4.5
Expected FE *e*	12.4	8.3	6.2	5.0	4.1	3.5	3.1	2.8	2.5	2.3

Table 3.1 shows the distance between categories, the maximum possible formatting error, and the mean expected formatting error with ordinal scales from three to twelve categories. The maximum formatting error occurs when the true value is exactly in the middle between two categories and the expected formatting error decreases with the number of categories available. The expected formatting error with 5-point scales is 6.2 percentage points and half as large with 9-point scales where the expected formatting error is 3.1 percentage points.

Formatting Error for Single Measurement with VASs. Regarding VASs, the results are unambiguous. Because this continuous rating scale offers a matching option for every

possible gradation of the latent interval scaled variable to be measured, the expected mean formatting error with VASs is zero.

3.2.2 Formatting error on the aggregate level

Even though it is illuminating to get a basic idea about the impact of the rating scale on the extent of formatting error, considering just a single measurement is not sufficient to estimate formatting error in real studies. In real tests or surveys one most often does not deal with a single measurement but with aggregate measurements. Data are either aggregated within respondents when indices are computed (e.g. when Likert's method of summed ratings is applied; Likert, 1932) or between respondents when means of variables or indices are computed. So, there is not just one true value as with a single measurement but there are up to as many true values as there are observations. The total formatting error depends on the actual distribution of these true values in a sample and it equals the sum of error occurring within each category. Figure 3.3 illustrates how *within-category error* is computed for a 5-point scale and a 6-point scale with normally distributed true values.

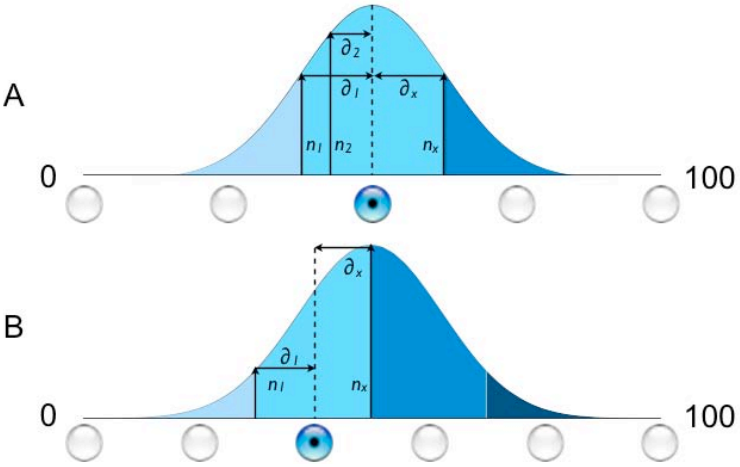


Figure 3.3. Within-category formatting error with normally distributed true values in 5-point scales (A) and 6-point scales (B).

The extent of within-category formatting error depends on two variables. It is influenced firstly by the difference δ between true value and chosen category value, and secondly by the number of cases with a certain difference. When assuming a normal distribution of true values and considering a 5-point scale relatively little formatting error occurs, as there are many cases that match the category value of the middle category. In comparison, with 4- and 6-point scales there are fewer cases without formatting error. Thus, formatting error is a gen-

eral argument against the use of even categories, because for normally distributed variables the lack of a middle category inevitably means formatting error occurring for a large number of cases. Figure 3.3 also illustrates that even slight differences from normal distributions can affect formatting error.

3.2.3 Method

To estimate the influence of formatting error on the interaction between the number of categories and differently distributed true values, we simulated data sets with different characteristics to model different populations. We generated random natural numbers between 0 and 100 following different distributions with different means and spreads. Random variables were generated using the open source software *R* (http://r-project.org; syntax files are available from the authors upon request). First, we generated 10^7 cases. We chose the central parameters so that most cases were in the range between 0 and 100. Then we excluded all values not falling between the boundaries 0 and 100 and finally drew a simple random sample of 10^6 cases to arrive at groups of equal size.

The first data set we simulated followed a uniform distribution where all values are about equally frequent. Secondly, we investigated normal distributions with different spreads and different means. We started with normal distributions with a mean in the exact middle of the continuum, at 50. As all scales with an odd number of response options share the category value 50, all odd scales should – depending on the spread of values – have a relatively low formatting error. To take this into account, we also investigated normal distributions with other means. We took 43 as the category value of the fourth category on an 8-point scale and 45 (the sixth option on a 12-point scale). Additionally, we simulated data with the mean on an odd category: 25 (second category on a 5-point scale).

Finally, there are some category values that odd and even scales share: 20 (being the second category on a 6-point scale as well as third on an 11-point scale), 33 (second option on a 4-point scale, the third option on a 7-point scale, and the fourth option on a 10-point scale), and 40 (representing the third option on a 6-point scale as well as the fourth option on an 11-point scale). It is obvious that the impact of the distribution on formatting error is more extreme when data have a small deviation from the mean. Thus, in order to illustrate extreme cases, we chose a (small) standard deviation of 5.0 for these shifted variables. In the end we simulated an exponential data set that should represent a very skewed distribution. All ten simulated data sets are illustrated in Figure 3.4.

Corresponding to the determination of formatting error for an individual measurement, we computed the distance to the next available response option for each possible true value in the data set for ordinal scales consisting of three to twelve categories. As above, the maximum formatting error equals about one fourth of the distance between categories. In practice, one would only arrive at this maximum in the unlikely case that all true values were at the border between two categories.

3.2.4 Results Study 1

Table 3.2 shows the aggregate formatting error for each of the ten simulated distributions. Only with a uniform distribution of true values (distribution 1) it equals the formatting error for a single measurement (leaving aside some small deviations because of rounding). Regarding the normal distributions with a mean of 50 (distributions 2 and 3), our prior reasoning shows to be correct: A large spread of values leads to formatting error very similar to error occurring with a uniform distribution. Overall, for a uniform distribution and normal distributions with a normal to wide spread of values, we find a confirmation for the initial reasoning: The larger the number of response categories the smaller the formatting error. This is also applicable for the exponential distribution we tested (distribution 10).

But when examining extreme distributions with small variances (distributions 4–9), we observe two exemptions from the rule mentioned above. First, formatting error does not always decrease with a larger number of categories. Second, an interaction between the number of categories and the distribution of values leads to a large spread of formatting error for response scales with a small number of categories. Even very small deviations of the mean from the middle of the continuum (e.g. a mean of 45 instead of 50) can have a large effect. With 3-point scales formatting error is greatly influenced by the distribution of values, ranging between 5.8 and 19.2 and with 4-point scales the expected formatting error is between 4.0 and 13.0. Even with 5-point scales error still is as high as between 3.9 and 8.1.

Taking these results together it becomes obvious that minimizing formatting error is only possible with knowledge about the exact distribution of parameters. Depending on the actual mean of a distribution, in some cases an odd number of categories leads to a smaller expected formatting error, in some cases an even number of response categories. Figure 3.4 illustrates the range of formatting error depending on the number of categories and the simulated distributions.

Table 3.2

Mean Formatting Error with Three to Twelve Categories and Differently Distributed Data

Distribution (D.)	Mean formatting error occurring with number of categories									
	3	4	5	6	7	8	9	10	11	12
1 Uniform	12.4	8.3	6.2	4.9	4.1	3.5	3.1	2.7	2.5	2.2
2 Normal D. (ND)										
(*M* = 50, *SD* = 15)	10.8	8.6	6.2	5.0	4.2	3.5	3.1	2.8	2.5	2.3
3 ND, wide										
(*M* = 50, *SD* = 45)	12.6	8.4	6.2	5.0	4.1	3.6	3.1	2.8	2.5	2.2
4 ND, narrow										
(*M* = 50, *SD* = 5)	**4.0**	13.0	**3.9**	6.2	**3.6**	3.7	**3.0**	3.1	2.5	2.5
5 ND, shifted 1, narrow										
(*M* = 45, *SD* = 5)	**5.8**	11.2	5.5	5.0	4.4	3.4	3.1	2.8	2.5	2.3
6 ND, shifted 2, narrow										
(*M* = 43, *SD* = 5)	**7.4**	9.7	6.7	**4.3**	4.8	3.3	3.1	2.8	2.5	2.3
7 ND, shifted 3, narrow										
(*M* = 40, *SD* = 5)	10.1	**7.3**	8.1	**3.8**	4.8	3.4	3.0	2.8	2.5	2.3
8 ND, shifted 4, narrow										
(*M* = 33, *SD* = 5)	16.8	**4.0**	7.2	5.7	3.6	3.6	3.3	2.7	2.5	2.2
9 ND, shifted 5, narrow										
(*M* = 20, *SD* = 5)	19.2	11.6	5.5	3.8	**3.8**	4.0	3.1	2.7	2.5	2.2
10 Exponential										
(*M* = 15)	10.3	7.5	5.9	4.8	4.1	3.5	3.1	2.7	2.5	2.2

Note. Values in bold face indicate occurrences where formatting error is *lower* in comparison to a rating scale with a *higher* number of categories.

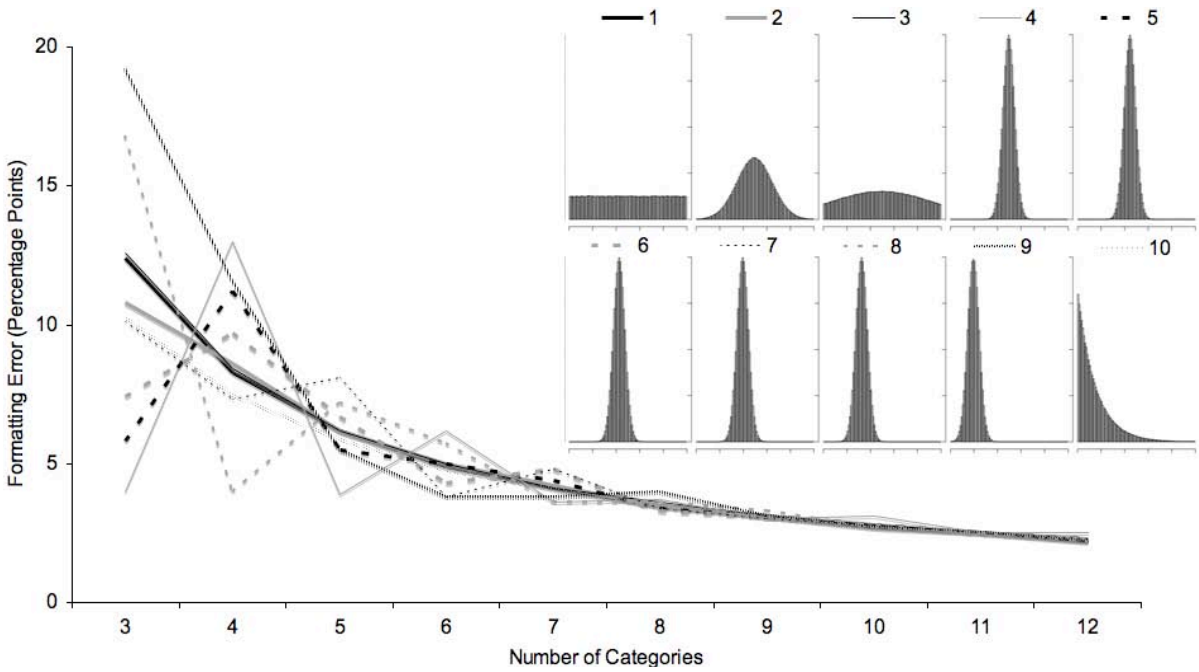


Figure 3.4. Formatting error for simulated data sets 1–10, as shown in Table 3.2.

With an increasing number of response options the impact of the distribution of values becomes smaller and smaller, as well as the absolute amount of formatting error. When con-

sidering scales with at least eight categories, from a practical perspective formatting error is quite similar for all distributions. However, even when using a large number of response categories the error affecting the measurement is still far from being zero. Using categorical scales for interval concepts thus adds noise to the data, and makes small effects unobservable in many designs. For measurement with VASs the same reasoning applies as above. As there is a matching value on the response scale for every true value possible the expected formatting error is zero.

To provide empirical evidence we conducted two independent Web experiments. We expect that measurements with VASs significantly outperform measurements with ordinal scales regarding measurement related error.

3.3 Study 2: Empirical error with 5-point scales and VASs

Study 2 was conducted to test if the theoretically lower formatting error with VASs is indeed observable in a field experimental study. Formatting error can only be measured directly if the true values are known, see e.g. Reips and Funke (2008) for an example. In all other cases it has to be inferred. One solution to compensate the availability of true values is to estimate formatting error indirectly by examining standard errors. A lower formatting error should find expression in a lower standard error of a measurement. Standard errors can easily be computed and lend themselves as indicator for the amount of error occurring during the measurement process. We expect measurement with VASs to lead to lower standard errors than measurement with ordinal rating scales. For analyses we used type of rating scale as independent variable and the mean standard error as dependent variable.

The standard error of the mean *SEM* is

$$SEM = \frac{s_M}{\sqrt{n}} \tag{2}$$

where s_M equals the standard deviation from the mean in the actual experimental condition and n is the number of observations in the sample. *SEM* becomes smaller with a larger number of observations. To take this into account, we drew a simple random sample from the VASs condition so that both conditions consisted of 230 cases.

3.3.1 Method

Procedure. In a Web experiment in a between-subjects design participants of a big five personality test (taken from the International Personality Item Pool IPIP[3.3], German version validated for Internet usage by Hartig, Jude, & Rauch, 2003) were randomly assigned to two conditions. The rating scales in the questionnaires were either made of 5-point ordinal scales or VASs of 250 pixels in length. Thus, respondents in the VASs condition had 50 times more possibilities to make their judgments. The first page of the questionnaire consisted of an introduction and a seriousness check (Reips, 2002c), followed by a short instruction that explained how to use the respective rating scale. Additionally, participants were asked for their gender and to provide an individual anonymous code in order to receive an individual feedback afterwards, if they so desired. The experimental section consisted of 40 Web pages in a one-item-one-screen design (see Reips, 2002c) and all items were voluntary to keep psychological reactance as low as possible (see Stiegler, Reips, & Voracek, 2007).

Participants. The participants were recruited in courses for first semester psychology students at the University of Kassel, Germany. Of the 585 participants who were eligible for the experiment, 505 (86.3%) indicated serious participation, whereof 246 (48.7%) were randomly assigned to the condition with 5-point scales and 259 (51.3%) to the VAS condition. 16 persons (6.5%) dropped out during the study in the 5-point scale condition and 19 persons (6.9%) with VASs, including one lurker[3.4]. Finally, 470 data sets remained for analyses, 230 cases (48.9%) with 5-point scales and 240 (51.1%) with VASs.

3.3.2 Results Study 2

As expected, the mean ratings of the 40 items were generally equivalent. We found three statistically significant differences with $p < .05$ (all eta^2 around .02). This general equivalence of measurement regarding mean ratings between ordinal scales and VASs was expected from the findings reported earlier.

As expected, in comparison to 5-point scales, *SEM* is lower for 27 variables (67.5%) with VASs, about the same for nine variables (22.5%), and slightly higher for only two variables (5.0%) in comparison to 5-point scales. On a scale from 0 to 100 the mean *SEM* with 5-point

[3.3] http://ipip.ori.org/ipip/
[3.4] Lurker are respondents who do not provide a single rating at all (see Bosnjak, 2001).

scales is M = 1.91 (SD = 0.38) and M = 1.69 (SD = 0.24) with VASs, $F(1,79) = 8.76$, $p = .004$, $\eta^2 = .10$.

In summary, we found evidence that VASs indeed have a beneficial effect on the reduction of error in the measurement process. A restriction is certainly that only 5-point scales were tested. Hence we set out to test other types of scales in Study 3.

3.4 Study 3: Formatting error with VASs versus 5-, 7-, and 9-point scales

To replicate and generalize the results from Study 2, we conducted a third study with a wider range of ordinal scales, and conducted it with a sample from outside the university. We compared VASs to three different ordinal scales: 5-point, 7-point and 9-point scales. Our first hypothesis is that a higher number of response options in ordinal scales leads to less error. The second hypothesis is that measurement with VASs should in any case be superior to measurement with ordinal scales, even with 9-point scales.

3.4.1 Method

Procedure. In a Web-based test of locus of control 58 items – presented in pairs on 29 Web pages – had to be evaluated. In the beginning of the questionnaire was a confidentially statement, a short introduction, a contact form, as well as a seriousness check. On the second page, participants were asked about their sex, age, formal education, and online experience. Afterwards, participants were randomly assigned to a condition where either 5-point scales, or 7-point scales, or 9-point scales, or VASs (349 pixels in length) were available.

Participants. The participants were recruited online in the Web Experimental Psychology Lab, via the Web Survey List, and in various newsgroups. Of the 585 eligible respondents 385 (65.8%) indicated serious participation. Table 3.3 illustrates the number of serious responds, respondents who dropped out of the study, and the respondents who completed the whole questionnaire. Two cases of lurking were observed with VASs. The final data set consisted of 265 cases.

Table 3.3

Serious Respondents, Dropout and Complete Cases by Experimental Condition

	5-Point scale	7-Point scale	9-Point scale	VAS
Serious (% between)	99 (25.7%)	111 (28.8%)	88 (22.9%)	87 (22.6%)
Dropout (% within)	42 (42.4%)	22 (19.8%)	33 (37.5%)	21 (24.1%)
Completes (% between)	57 (21.5%)	89 (33.6%)	55 (20.8%)	64 (24.2%)

3.4.2 Results Study 3

We compared the mean ratings of the 58 substantial items between ordinal scales and VASs. A statistically significant difference ($p < .05$) was found between VASs and 5-point scales with five variables (η^2 between .03 and .08), between VASs and 7-point scales with one variable ($\eta^2 = .03$), and between VASs and 9-point scales with five variables (η^2 between .03 and .04). The difference in dropout was statistically significant between VASs and 5-point scales, $\chi^2(1, N = 186) = 8.05$, $p = .005$, as well as between VASs and 9-point scales, $\chi^2(1, N = 174) = 4.16$, $p = .049$, but not between VASs and 7-point scales, $\chi^2(1, N = 197) = 0.34$, $p = .601$.

As in Study 2 we adjusted the number of cases for every experimental condition by drawing a single random sample, leading to 55 cases in every group. Afterwards, we compared *SEM* between VASs and each ordinal scale by using the rating scale as independent variable and the standard error as dependent variable. Table 3.4 shows how *SEM* differed between ordinal scales and VASs.

Table 3.4

Number of Items with Higher, Lower or Equal SEM Measured with Ordinal Scales in Comparison to Measurement with VASs

		5-Point scale	7-Point scale	9-Point scale
SEM	lower than VASs	7 (12.1%)	11 (19.0%)	20 (34.5%)
	equal to VASs	4 (6.9%)	12 (20.7%)	3 (5.2%)
	lower with VASs	47 (81.0%)	35 (60.3%)	35 (60.3%)

Mean *SEM* was $M = 3.86$ ($SD = 0.49$) with 5-point scales, $M = 3.69$ ($SD = 0.39$) with 7-point scales, $M = 3.56$ ($SD = 0.41$) with 9-point scales, and $M = 3.47$ ($SD = 0.40$) with VASs. The difference in *SEM* between VASs and 5-point scales is statistically significant, $F(1, 115) =$

21.55, $p < .001$, $\eta^2 = .16$, as well as the difference between VASs and 7-point scales, $F(1, 115) = 8.95$, $p = .003$, $\eta^2 = .07$. The difference between VASs and 9-point scales did not reach statistical significance, $F(1, 115) = 1.56$, $p = .215$.

3.5 Discussion

It is everyday practice to use ordinal rating scales to measure continuous variables. Our analysis and evidence from three studies show: this way of collecting data unavoidably raises the total survey error by producing formatting error. In Study 1 we found that for ordinal scales up to about seven categories there is a considerable interaction between the number of response options and the distribution of true values, making the actual amount of error produced by the rating scale unpredictable. But even if more response options are provided, ordinal scales still add unnecessary error. Error widens confidence intervals, leads to less statistical power, makes it impossible to detect small effect sizes, and raises the risk of making a Type II error. All these undesirable effects can be avoided when using VASs, adequate rating scales that measure continuous data on the level of an interval scale.

Study 2 and Study 3 provide empirical evidence to the reasoning that measurements with VASs contain less error. First, we showed there is little foundation for the concern that VASs could bias data or raise dropout. The means of the items neither differed to a great extent nor systematically. Dropout with VASs is even lower than dropout with most ordinal rating scales, and thus indicates that respondents do not perceive using this rating scale as burdensome. Second, the amount of error in the measurement process decreases with an increasing number of response options. A lower error was observed with VASs in comparison to ordinal rating scales with up to seven response options.

With the proliferation of computer-based surveying and the advent of Web applications for Internet-based research, the creation and use of VASs has become easy and practical (Reips & Funke, 2008). Given the theoretical advantages and empirical results from the present series of studies, the benefits of VASs for data collection should be taken advantage of. The fine gradation of the response scale allows the detection of small differences between and within subjects. Additionally, data obtained with VASs can be recoded into any number of categories without losing the desired scale properties. If data are transformed to equidistant categories consisting of equal intervals, the conditions for the usage of statistical methods that require data on the level of an interval scale are still met. When the usage of VASs

is not possible due to technological considerations or for practical reasons, the authors suggest the use of rating scales with more than seven response options.

Overall, the use of ordinal rating scales to measure continuous variables is an inappropriate methodology. The authors strongly advocate to reconsider this frequently observed practice and instead make use of the advantages of VASs for the measurement of continuous variables.

4

Semantic Differentials Made of Visual Analogue Scales in Web-Based Reserach

4 Semantic Differentials Made From Visual Analogue Scales in Web-Based Research[4.1]

In a Web experiment, participants were randomly assigned to 2 semantic differentials either made from discrete 5-point ordinal rating scales or continuous visual analogue scales (VASs). Respondents adjusted their ratings with VASs more often to maximize the precision of answers, which had a beneficial effect on data quality. No side effects like higher dropout, more nonresponse or higher response times were observed. Overall, the combination of semantic differentials and VASs indicates a number of advantages. Potential for further research is discussed.

4.1 Semantic differentials made from visual analogue scales[4.1]

Research on the historic origin of semantic differentials revealed that this type of rating scale was initially made from continuous, not from discrete rating scales. McReynolds and Ludwig (1987) report that a device very similar to what we consider a semantic differential nowadays – even though it did not contain contrary verbal labels on either side – was used as early as the beginning of the 19th century: "a metal plate […] had 10 scales, each marked off in 100 parts, and labeled […]. A system of sliding markers was provided so that a […] judged position on each scale could be graphically displayed" (p. 282). More than 180 years later we transfer this approach into cyberspace, substituting the metal plate with a Web browser's interface.

4.2 Introduction

Researchers have an extensive methodological repertoire at their hands to design Web-based questionnaires that assist participants in giving accurate answers and maintaining their willingness to cooperate. One possibility to optimize questionnaire design is to alter the

[4.1]Parts of this study have been presented at the ISA RC33 7th International Conference on Social Science Methodology, Naples, Italy, September 1–5, 2008, under the title "Assessing semantic differentials with visual analogue scales in Web surveys". The authors would like to thank the anonymous reviewers for their helpful remarks as well as E.-D. Lantermann (University of Kassel) and his students who participated in the experiment. This chapter has been submitted for publication:
Funke, F., & Reips, U.-D (2010). *From the roots to the future: Semantic differentials made from visual analogue scales in Web-based research.* Manuscript submitted for publication.

available response scales. Changes in response scales can affect the question answer process, especially question understanding as well as the formatting of answers (e.g. Sudman, Bradburn, & Schwarz, 1996), and seriously impacting given ratings (see deLeeuw, Hox, & Dillman, 2008; Dillman, Smyth, & Christian, 2009; Funke, Reips, & Thomas, in press; Groves, Fowler, Couper, Lepkowski, Singer, & Couper, 2004; Krosnick, 1999; Schwarz, 1999). The number of response categories communicates how elaborated the expected answer should be. A small number of response options implicitly conveys the message that roughly estimated answers are sufficient, whereas a large number of response options can be understood as an instruction to maximize cognitive efforts (see also Schwarz, 1999).

4.2.1 Semantic differentials

C. E. Osgood is credited for introducing semantic differentials in the 1950s (Osgood, 1952; Osgood, Suci, & Tannenbaum, 1957). By now it is an established measurement device used in many fields (e.g. psychology, sociology, and linguistics). Semantic differentials are used to assess different dimensions of a single construct. A battery of contrasting bipolar verbal anchors (e.g. *warm – cold* or *bright – dark*) is presented in form of a matrix.

It is known that respondents are likely to see items not independently but in the context of one another (see Dillman et al., 2009). Thus, when measuring unrelated items it is advantageous to present each item on a separate page (see Reips, 2002c) to minimize anchoring and contrast effects (e.g. Sudman, Bradburn, & Schwarz, 1996). In a semantic differential all dimensions are intentionally presented on the same (Web) page to emphasize that all dimensions are mutually related and that respondents are asked to give sound ratings regarding all dimensions.

4.2.2 Web-based data collection

Web-based data collection has become a well-established instrument in the world of survey methodology (Best & Krueger, 2004; deLeeuw, Hox, & Dillman, 2008; Dillman & Bowker, 2001; Dillman et al., 2009; Joinson, McKenna, Postmes, & Reips, 2007). In comparison to laboratory settings, Web-based research for example greatly extends the access to large number of participants and to special populations (see Fuchs, 2008; Mangan & Reips, 2007), and it allows the presentation of various multimedia stimuli (e.g., Fuchs & Funke, 2007).

Nevertheless, there are some pitfalls. For example, respondents have to have a certain degree of computer literacy to be able to complete a Web-based questionnaire in a meaningful way. Another worry is that the setting in which respondents take part in a study cannot be controlled. One the one hand, these non-standardized situations can be a good reason why Web experiments sometimes may not produce the same results as experiments conducted in laboratories (Reips, 2002c, 2007). One the other hand, the validity of results obtained from testing persons in real-life environments without (involuntary) influence of the experimenter should be higher (see Honing & Reips, 2008).

Especially Web-based questionnaires are prone to involuntary changes in layout due to different client-sided software configurations. For example poor HTML code can result in uneven spacing between radio buttons, which can shift ratings as Tourangeau, Couper, and Conrad (2004) demonstrated. As even small changes in layout and visual design can affect answers (e.g. Couper, Traugott, & Lamias, 2001; Dillman & Bowker, 2001; Reips, 2010; Smyth, Dillman, Christian, & Stern, 2006), one should pay much attention to a straightforward (e.g. low-tech) implementation and control for technological background variables (e.g. operating system, browser, and availability of technologies).

4.2.3 Visual analogue scales (VASs)

VASs are continuous graphical rating scales, first described by Hayes and Paterson (1921). The obvious advantage over discrete scales is that answers are not restricted to a certain number of response options but very fine gradations can be measured. In computerized data collection each pixel in length corresponds to a possible value (Reips & Funke, 2008).

Many paper-based studies – especially in the medical sector, e.g. for assessment of subjective phenomena like fatigue or pain (Cork, Isaac, Elsharydah, Saleemi, Zaviska, & Alexander, 2004) – were not able to show differences between VASs and ordinal scales regarding mean ratings (also see Averbuch & Katzper, 2004; Flynn, van Schaik, & van Wersh, 2004). In a paper and pencil study on pain, Myles, Troedel, Boquest, and Reeves (1999) and Myles and Urquhart (2005) found that data from VASs are linear. Gerich (2007) found that mode differences between paper and Web are smaller with VASs than with 5-point scales. Hofmans and Theuns (2008) conclude that in a paper-based study “VASs can be considered as linear scales and that the type of end anchors used has no effect on the linearity of the VAS data” (p. 401). Reips and Funke (2008) found that even in Web surveys VASs fulfill many requirements of measurement on the level of an interval scale. With mentally well-represented con-

structs equal changes in intensity correspond to equal changes in ratings on VASs. So, differences between ratings on VASs can be interpreted in a meaningful way and the prerequisites for many statistical procedures are met.

In contrast to the historical origins, semantic differentials are rarely made of VASs nowadays. The most likely explanation is the effort associated with the manual readout in paper-based studies. However, this is no issue in computerized data collection.

4.3 Research questions and hypotheses

Making ratings on semantic differentials is a more complex task than just answering single items. Respondents have to consider not only a single item but also the relationship between each item presented to give sound overall judgments. Our research focuses on how the rating scale used with semantic differential influences the measurement process and data quality.

VASs allow giving finely nuanced ratings regarding all dimensions of the construct being measured. However, this large number of possibilities may be too demanding. Ratings with *ordinal scales* pose the problem of shared ranks whenever the number of response options is smaller than the number of items. Thus, giving sound ratings regarding the relation between all items is impossible. Ordinal rating scales always require decisions about compromises and thus may increase cognitive burden. Additionally, if continuous concepts are to be rated on ordinal scales, a transformation, a segmentation and mapping to the nearest category have to be mastered by the respondent.

Web surveys allow unobtrusive data collection with nonreactive methods (e.g. Heerwegh, 2003). Using JavaScript we implemented a measure of how long it takes to make ratings and how often ratings are modified. Thus, we made use of the enhanced technologies available in Internet-based research when investigating the manual answering process in remote participants (Stieger & Reips, in press). Regarding the observable part of the question-answer process, we hypothesize that respondents take advantage of the possibilities of VASs and change given ratings more frequently in comparison to ordinal scales. In turn, this should raise response times and lead to higher data quality.

4.4 The experiment

The Web experiment was conducted as part of a study on personal preferences, taste, and style preferences among students of psychology at the University of Kassel, Germany, from January 7–30, 2008. Respondents provided a personal, but anonymous code to be able to get an individual feedback as per capita incentive (see Göritz, 2006).

4.4.1 Questionnaire

The questionnaire consisted of 83 consecutive Web pages. To utilize advantages of the warm-up technique in Internet-based research (Reips, 2002c) the experimental manipulation of the type of rating scales in the semantic differentials was placed on pages 81 and 82. The questionnaire contained no other semantic differentials. On each of these two pages was one semantic differential consisting of 13 items, assessing different dimensions of the respondents' style regarding furnishing (semantic differential 1) and clothing (semantic differential 2). The labeling of the 13 dimensions was the same on both pages.

4.4.2 Procedure

In a between-subjects design respondents were randomly assigned either to semantic differentials made from 5-point rating scales, implemented with HTML radio buttons, or to semantic differentials made from VASs with 250 pixels in length corresponding to 250 gradations (see Figure 4.1). The VASs were generated with the free Web service VAS Generator (maintained by the authors, located at http://vasgenerator.net) and implemented using JavaScript. Type of rating scale was the same on both Web pages.

Directly after loading the Web pages containing the semantic differentials, neither scale showed any marker (Figure 4.1, top six items). Indicators for previously given ratings – a checked radio button with the 5-point scales and a cross with the VASs (see Figure 4.1, bottom items) – only appeared after clicking on the scales. Judgments with VASs were made in the following way (working examples can be seen at http://vasgenerator.net): Respondents clicked the blank line at the appropriate position and the marker appeared at the very position. To modify the rating any other position on the VAS could be clicked. It was not possible to move the marker by dragging it with the mouse. Every click on the VASs was recorded just

as every click on a radio button. In both conditions respondents could adjust all ratings as often as they wanted and no rating was mandatory.

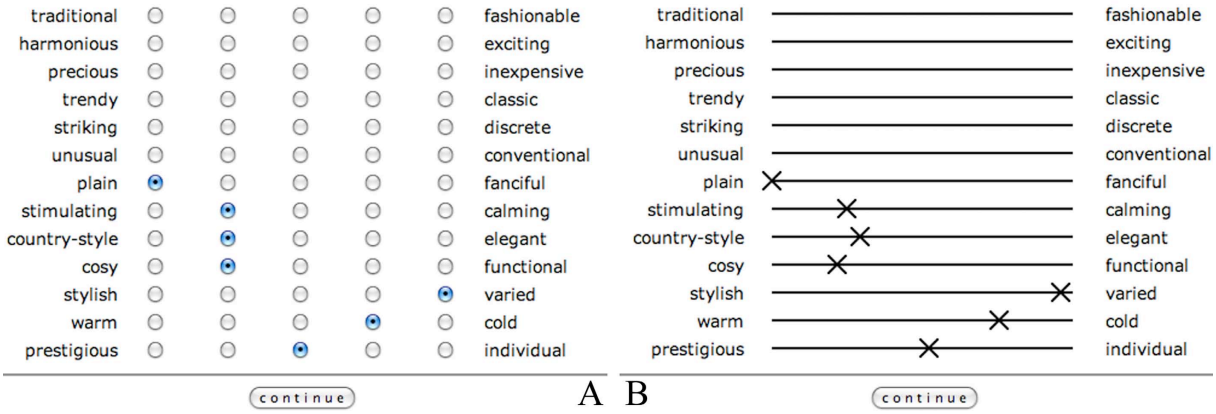


Figure 4.1. Ratings scales (top after load bottom with ratings): radio buttons, five categories (A) and VASs, 250 pixels (B).

4.5 Results

4.5.1 Participants

Overall, 278 participants reached the first page of the experiment. At the beginning of the survey, participants were asked for an anonymous code. Not providing a code indicated a low motivation to seriously participate in this study, so we excluded these seven cases. We excluded three additional cases (1.1%) of JavaScript not being enabled in the respondents' Web browsers. Overall, we thus obtained a sample of 268 cases, 134 in each experimental condition. Due to the warm-up technique (Reips, 2002c) no *dropout* occurred on the two pages of the experiment.

4.5.2 Nonresponse

In both conditions 7.1% of the data set (19 cases with each scale) were incomplete, showing at least one missing value during the experiment. Following the common definition of *lurking* (e.g. Bosnjak, 2001), we take lurkers as respondents who did not provide even a single answer within a survey. We detected two lurkers, one per condition. Additionally, we found two cases with ordinal scale type that did not provide a single value for either semantic differential (one-page lurker). We excluded these four cases, resulting in 264 cases (133 with VASs and 131 with 5-point scale) for analyses of item nonresponse.

We found no statistically significant difference in item nonresponse between VASs and the 5-point scales. However, there was a trend that is consistent within both semantic differentials: Semantic differential 1: *M*(VAS) = 0.28 (*SD* = 1.34), *M*(5-point) = 0.42 (*SD* = 1.63); semantic differential 2: *M*(VAS) = 0.40 (*SD* = 1.72); *M*(5-point) = 0.50 (*SD* = 1.89).

Prior to analyzing frequency of changing answers, means, correlations, and response time, we additionally excluded all cases with at least one missing item, resulting in a net sample of 230 cases, 115 with each type of scales. All in all, we excluded 17.3% of respondents from analyses. The female proportion in the net sample was 76.5% with VASs and 71.3% with 5-point scales. The mean age was *M* = 23.1 years (*SD* = 2.2) with VASs and *M* = 23.7 (*SD* = 3.8) years with 5-point scales. None of these demographic differences between the experimental conditions was statistically significant (all $p > .10$).

4.5.3 Adjusting responses

To make inferences on the process of decision making we analyzed how often ratings were modified. The effort to give accurate responses is reflected in the number of changes. To be more conservative, we looked for extreme values of changes and disregarded data from one respondent (53 changes with VASs and the second semantic differential). Overall, with the first semantic differential 66.1% of the respondents changed at least one answer with VASs and 57.4% with the 5-point scale. Fewer changes were made with the second semantic differential: 54.8% with VASs and 50.4% with the 5-point scale.

On the page with the first semantic differential, the mean number of clicks needed to complete all 13 items was *M* = 16.7 (*SD* = 5.0) with the VASs and *M* = 14.5 (*SD* = 2.1) with the 5-point scales, $F(1, 229) = 19.50$, $p < .001$, $\text{eta}^2 = .079$. In other words: with VASs 3.7 changes occurred with the first semantic differential and only 1.5 with the 5-point scales. On page two, answering the semantic differential with VASs, *M* = 14.9 (*SD* = 2.8), again took more clicks than answering with the 5-point scales, *M* = 14.0 (*SD* = 1.7), $F(1, 228) = 8.19$, $p = .005$, $\text{eta}^2 = .035$. This corresponds to 1.9 changes with VASs and 1.0 change with the 5-point scales.

4.5.4 Ratings

For better comparability, data from both scales were recoded ranging from 0 to 100. In the first semantic differential (see Figure 4.2), only the mean rating for the dimension *distin-*

guished – individual differed significantly, *M*(VAS) = 67.7 (*SD* = 23.3), *M*(5-point) = 60.9 (*SD* = 25.9), *F*(1, 229) = 4.47, *p* = .036, eta^2 = .019.

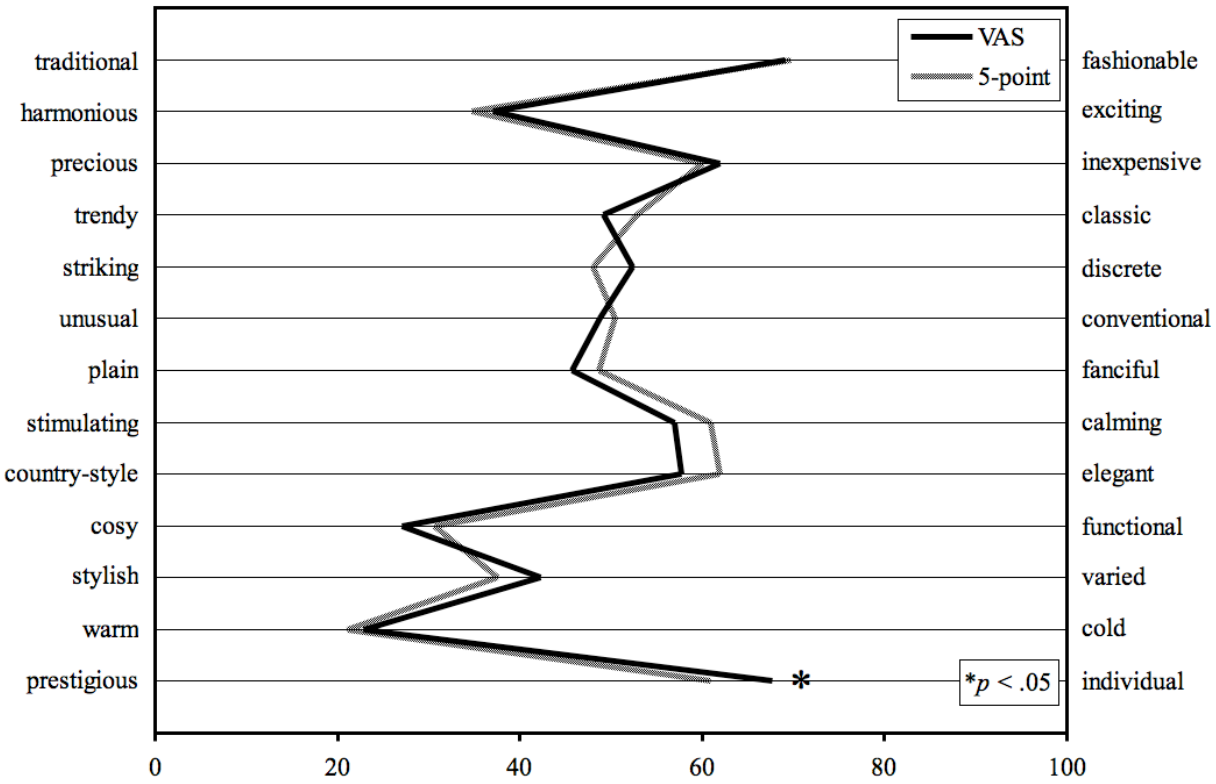


Figure 4.2. Mean ratings semantic differential 1.

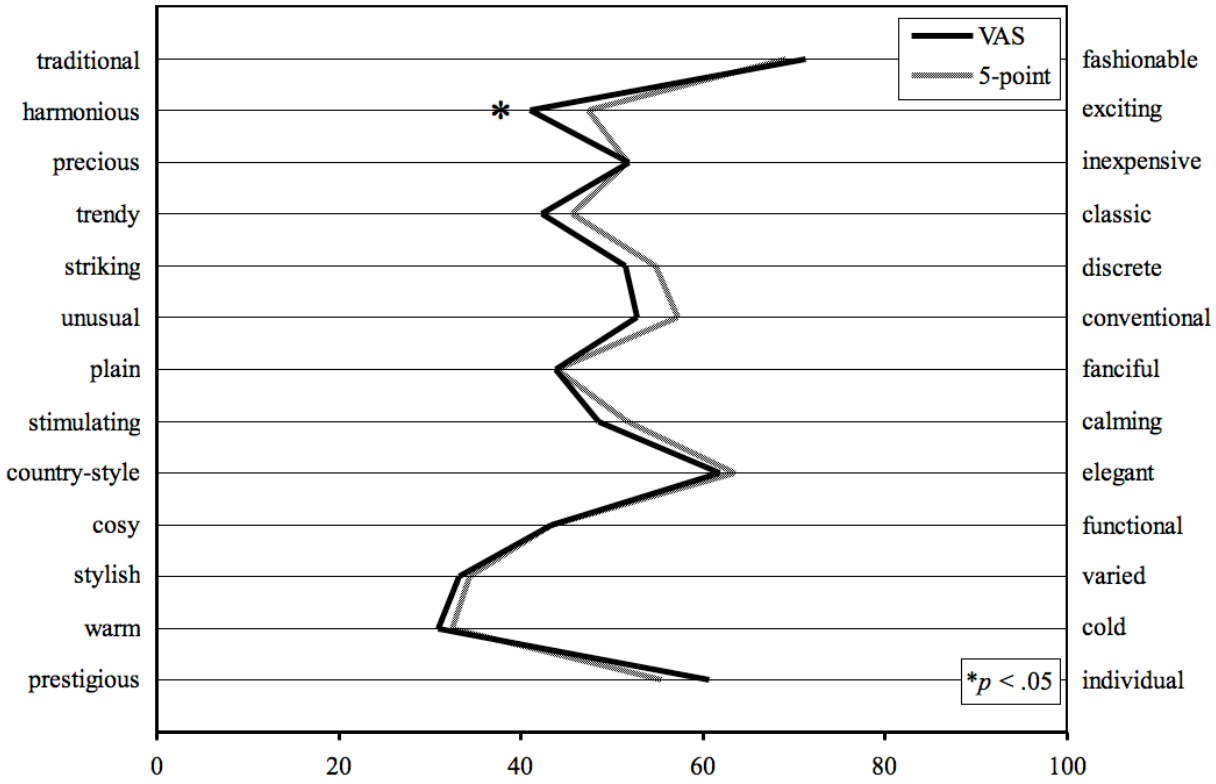


Figure 4.3. Mean ratings semantic differential 2.

With the second semantic differential, the picture is quite the same (see Figure 4.3). Again, just one dimension (*harmonious* – *exciting*) showed a statistically significant difference in means: *M*(VAS) = 41.0 (*SD* = 23.6), *M*(5-point) = 47.4 (*SD* = 24.2), *F*(1, 229) = 4.17, *p* = .042, eta^2 = .018.

4.5.5 Correlations

Both concepts we observed with the semantic differentials (style of furnishing and style of clothing) should be highly correlated as we expect a general factor "sense of style" to influence both variables. We used the correlations between identical dimensions (e.g. dimension *plain* – *fanciful* for furniture style and *plain* – *fanciful* for clothing style) to infer on measurement error. A desired measurement with lower error would show high correlations between variables measuring the same dimension and low correlations – indicating high discrimination – between different dimensions (e.g. *plain* – *fanciful* and *trendy* – *classic*). We computed Pearson's correlation coefficient between all 13 variables for comparing measurement with VASs to measurement with 5-point scales (see Table 4.1).

Correlations Between Corresponding Dimensions. For this analysis we looked at the correlations between the 13 corresponding dimensions (see Table 4.1, printed in bold type). Mean correlations measured with VASs (*M* = .48, *SD* = .10) were higher than correlations with the 5-point scales (*M* = .41, *SD* = .11). This difference is marginally significant and has a large effect size, *F*(1, 25) = 3.41, *p* = .077, η^2 =.125.

Correlations Between Distinct Dimensions. If VASs produced higher correlations in general, not only with corresponding dimensions but also with unrelated dimensions, this would be an undesired effect (e.g. always just clicking in the middle of a rating scale would have produced high correlations between all dimensions). Thus, we computed the mean correlations for all unrelated dimensions (see Table 4.1, printed in regular type). Mean correlations between unrelated dimensions of the construct *style* were zero measured with VASs (*M* = .00, *SD* = .21) and nearly zero with the 5-point scales (*M* = -.01, *SD* = .17).

Table 4.1

Correlations Between Dimensions for VASs and 5-Point Scales

Dimension	1	2	3	4	5	6	7	8	9	10	11	12	13
	VASs (*n* = 115)												
1	**.47****	.07	-.23*	-.20*	-.31**	-.20*	.18	-.15	.25**	.08	-.23**	-.11	-.05
2		**.43****	.18	.13	-.19*	-.10	.06	-.21*	-.29**	.02	.46**	.16	.04
3			**.63****	.10	.36**	.18	-.30**	.26**	-.35**	-.25**	.35**	-.09	.09
4				**.46****	.39**	.38**	-.32**	.26**	-.02	-.10	.19*	.20*	-.31**
5					**.46****	.36**	-.36**	.33**	.19*	-.01	-.17	.14	-.04
6						**.54****	-.36**	.24**	.18	-.14	-.21*	.01	-.18
7							**.60****	-.18*	-.11	-.09	.11	-.14	.19*
8								**.36****	.07	.01	-.15	.05	.08
9									**.56****	.00	-.46**	-.20*	.00
10										**.34****	.02	.38**	-.10
11											**.64****	.15	.02
12												**.40****	-.01
13													**.39****
	5-point scales (*n* = 115)												
1	**.46****	.25**	-.30**	-.37**	-.27**	-.11	.20*	-.11	.22*	.13	-.15	.00	-.04
2		**.35****	.01	-.11	-.01	-.15	-.07	.00	-.16	.20*	.20*	.38**	.08
3			**.44****	.26**	.09	-.03	-.16	.05	-.34**	-.06	.32**	.05	.11
4				**.44****	.18	.14	-.32**	.13	-.18	-.02	.09	-.08	.05
5					**.33****	.24*	-.35**	.21*	-.02	-.15	.07	.01	-.05
6						**.46****	-.24**	.16	.10	-.13	-.15	-.10	-.21*
7							**.48****	-.21*	.06	-.08	-.03	-.12	-.01
8								**.17****	.04	-.25**	-.01	.02	-.05
9									**.58****	.12	-.26**	-.06	-.14
10										**.33****	-.08	.28**	.00
11											**.53****	.18	.14
12												**.39****	.17
13													**.32****

Note. Correlations between corresponding dimensions are in bold type.
*Correlation is significant at the .05 level (two-tailed). **Correlation is significant at the .01 level (two-tailed).

4.5.6 Response times

In contrast to a laboratory setting, there can be multiple sources of participant distraction in Web surveys not related to the study (e.g. incoming phone calls). We decided to remove all

unreasonably high response times from analyses and identified outlier as proposed by Tukey (1977): Within every experimental condition, all response times lower than 25th percentile minus one and a half times the interquartile range were omitted as well as times higher than 75th percentile plus one and a half times the interquartile range. For further analyses only respondents with moderate response times on both pages were considered, resulting in 33 respondents (15 with VASs and 18 with 5-point scales) that were excluded from analyses of response time.

With the first semantic differential completion with VASs (*M* = 66.4 seconds, *SD* = 19.4) did not take statistically longer than completion with 5-point scales (*M* = 62.7 seconds, *SD* = 20.0). On the second page, the absolute difference was even lower: *M*(VAS) = 48.9 seconds (*SD* = 13.9), *M*(5-point) = 47.8 seconds (*SD* = 14.1).

4.6 Discussion

VASs allow respondents to communicate subjective values more exactly than radio button scales. This advantage should be especially valuable with semantic differentials, the prime method for assessing multiple dimensions of one construct on a single (Web) page. Aim of this study was to test if the theoretical advantages of semantic differentials made from VASs in comparison to semantic differentials made from 5-point scales hold in a Web experiment.

4.6.1 Decision making processes

We considered the response times and frequency of adjusting responses to infer on decision making processes. VASs had a clear positive influence on the number of changes. Ratings with VASs were modified around twice as often as with 5-point scales. In line with our hypothesis, respondents indeed made use of VASs' fine gradations. Are more adjustments with VASs an indicator for deeper cognitive processing? In contrast to our expectations, we found no difference in response times. Regardless of the available rating scale respondents seem to be willing to invest a certain amount of time for dealing with the task of answering the items in a semantic differential. With discrete 5-point scales' limited number of response options and the problem of shared ranks this time is likely to be used for *formatting the answer,* i.e. to find the best fitting response option. In contrast, with continuous VASs respondents do not have to bother about restrictions of the rating scale. Instead they use the time to *maxi-*

mize the precision of the given answers. This process of maximizing efforts should find expression in data quality.

The indicator for data quality we used was the correlation between style of furnishing and style of clothing. Our reasoning is that both domains of style should be highly correlated as influenced by the same general factor. A superior measurement should lead to high correlations between corresponding dimensions of style and to no correlations between distinct dimensions. Indeed, correlations between corresponding dimensions were significantly higher with VASs in comparison to 5-point scales. Correlations between distinct dimensions are around zero for both scales. We take these findings as indicator that measurement with VASs has a beneficial influence on data quality.

4.6.2 Mean ratings and non-response

Overall, mean ratings are hardly affected by the rating scales we tested, the absolute difference was very small. This is in line with a large body of literature on VASs yielding the same mean score as ordinal scales (e.g. Averbuch & Katzper, 2004; Flynn et al., 2004; for Web-based research see Couper et al., 2006; Funke & Reips, 2010). We found statistically significant differences only for two out of 26 items, which can be attributed to chance. Additionally, effect size was small, showing that the type of rating scale explains less than two percent of the variance (for comparison: we found a statistically significant effect of gender on twelve items explaining between three percent and ten percent of the variance).

Regarding general indicators of data quality – dropout, lurking and item nonresponse – our results show no statistically significant differences between rating scales. There is – in contrast to Couper et al. (2006) – a tendency for less item nonresponse with VASs, which may be owed to the low-tech implementation of VASs in the study presented.

4.7 Conclusion

For the first time, the present study compared semantic differentials made up from either VASs or ordinal scales. VASs are suited to (literally) draw an accurate picture of all dimensions of the construct being measured. In contrast, ratings with 5-point scales are troubled by shared ranks and the low degree of differentiation between dimensions. We consider the results reported as encouraging for using VASs with semantic differentials. Further research

with eye tracking (e.g. Galesic, Tourangeau, Couper, & Conrad, 2008) or the User Action Tracer technique (Stieger & Reips, in press) could help understanding the exact way ratings are made with semantic differentials. Our recommendation is to go back to the historical roots and to use VASs for Web-based data collection with semantic differentials.

5

Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level

5 Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level[5.1]

Slider scales and radio buttons scales were experimentally evaluated in horizontal and vertical orientation. Slider scales lead to statistically significantly higher break-off rates (odds ratio = 6.9) and substantially higher response times. Problems with slider scales were especially prevalent in participants with less than average education, suggesting the slider scale format is more challenging in terms of previous knowledge needed or cognitive load. An alternative explanation, *technology-dependent sampling* (Buchanan & Reips, 2001), cannot fully account for the present results. The authors clearly advise against the use of Java-based slider scales and advocate low-tech solutions for the design of Web-based data collection. Orientation on screen had no observable effect on data quality or usability of rating scales. Implications of item format for Web-based surveys are discussed.

5.1 Introduction

Participation in a survey is a special kind of indirect communication between respondent and researcher. As in self-administered studies no interviewer is present to provide situational clarification aids, respondents use various elements in a questionnaire as cues to infer further information from. In turn, questionnaire design elements and rating scales can be used to help respondents to understand a question in a certain desired way, reducing the variance of perceived question meaning and increasing the data quality. In this paper we focus on *rating scales* and *spatial orientation* as a means of design in a Web-based questionnaire (e.g. Couper, 2008; Dillman, Smyth, & Christian, 2009; Reips, 2000), and how the design may interact with respondent variables such as education (Buchanan & Reips, 2001).

5.1.1 Design effects and Web-based research

Questionnaire layout and graphical design (for an overview see Dillman et al, 2009; Lyberg et al., 1997) can have a considerable impact on data quality and substantive answers in both paper-based questionnaires and Web surveys (e.g. Christian, Dillman, & Smyth, 2005; Cou-

[5.1]This chapter is based on a presentation given at the 36th Annual Meeting of the Society for Computers in Psychology, Houston, TX, on November 16, 2006. It has been accepted for publication as: Funke, F., Reips, U.-D., & Thomas, R. K. (in press). Sliders for the smart: Type of rating scale on the Web interacts with educational level. *Social Science Computer Review.*

per, Tourangeau, Conrad, & Singer, 2006; Couper, Traugott, & Lamias, 2001; Reips, 2010; Tourangeau, Couper, & Conrad, 2004, 2007). Rating scales can affect data quality regarding mean ratings, distribution of answers, response time, or item nonresponse (e.g. Couper, Conrad, & Tourangeau, 2007; Healey, 2007; Heerwegh & Loosveldt, 2002; Krosnick, 1999; Krosnick & Fabrigar, 1997; but see Reips, 2002a). Especially Web-based questionnaires are prone to involuntary changes in design (e.g. leading to a nonlinear format, see Christian, Parsons, & Dillman, 2009). These changes mainly occur because of a considerable amount of variation of different hardware and software on the developers' and the respondents' side (Schmidt, 2007). Consequently, a robust low-tech implementation – where serious differences are less frequent – is the method of choice (Reips, 2002b, 2006).

5.2 Experimental manipulation

To gain more information about how rating scales and their design influence data and data quality, we conducted a 2 x 2 Web experiment (see Reips, 2007). Firstly, we manipulated the *spatial orientation* of the rating scale on the screen. Secondly, we tested two different *rating scales*, a slider scale and a common radio button scale (see Figure 5.1). Furthermore, we looked at high versus low level of education.

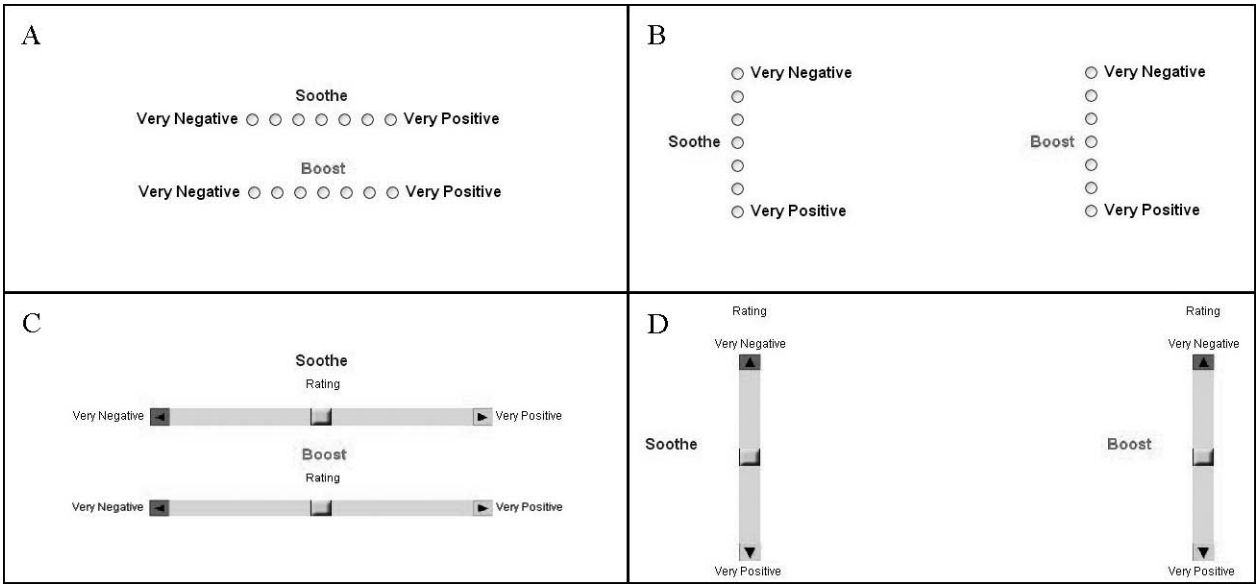


Figure 5.1. Rating scales used in the experimental conditions in Study 1: radio buttons (A and B) and slider scales (C and D) in horizontal (A and C) and vertical (B and D) orientation.

5.2.1 Factor 1: Horizontal versus vertical orientation on screen

Graham (1956) compared horizontal and vertical rating scales and found "shape of the visual field and the relative ease of moving the eyes from side to side, rather than up and down, are thought to account for the greater accuracy on the horizontal scale" (p. 157). Changes in the orientation of graphical rating scales were examined by Downie et al. (1978) as well as by Scott and Huskisson (1979) in paper-based questionnaires. Both compared horizontal and vertical visual analogue scales – graphical rating scales that are somewhat similar to slider scales – in within-subjects designs and found high correlations between both scales.

There is only little research in Web-based studies dealing with vertical or horizontal design, but it is mostly limited to close-ended questions (e.g. Smyth, Dillman, Christian, & Stern, 2006) or fully-anchored rating scales rather than end-anchored scales. Vertical rating scales may be especially interesting for mobile Web surveys (see Fuchs, 2008), as most mobile Internet devices (e.g. smart phones) have an upright display that may not display horizontal scales without scrolling. Vertical rating scales on displays in horizontal orientation (and vice versa) could bias results, as response options that are directly visible are preferred to response options only visible after scrolling (Couper, Tourangeau, Conrad, & Crawford, 2004).

5.2.2 Factor 2: Slider scale versus radio button scale

Levels were either a simple ordinal rating scale, realized with HTML radio buttons, or a slider scale programmed in Java. Both scales provided seven discrete values.

Slider scales are graphical rating scales not available for self-administered paper and pencil interviews because a slider has flexible parts that cannot be printed in a questionnaire. In our case the slider handle was initially positioned in the middle of the rating scale. In contrast to radio buttons scales – on which the appropriate answer is chosen by a single click – ratings with slider scales may be more demanding and require more hand-eye coordination: Respondents have to move the mouse pointer to the slider, then click and hold the mouse button, move the slider to the desired position, and release the mouse button.

An inherent problem with slider scales is the initial positioning of the handle, which could influence the probability of choosing the pre-selected value, resulting in biased estimates (e.g. if respondents interpret it as a typical judgment or the most desired answer). Even if the handle is initially positioned outside the scale anchoring effects could occur. Additionally, if the

handle is placed at the position of a valid answer option attitude and non-attitude cannot be distinguished.

Slider scales cannot be programmed in plain HMTL, more sophisticated programming (e.g. Java, JavaScript, Flash) is necessary. Client-side availability of the respective technology determines if a slider scale can be used or not. If a technology is not available in the respondent's browser it can result in item nonresponse, transmission of a default or invalid value, or even in respondents dropping out of a study (Buchanan & Reips, 2001).

Arnau, Thompson, and Cook (2001) compared an unnumbered slider scale with a 9- point radio button scale. The slider scale was programmed in Java like the slider used in this study, but in contrast it provided 100 discrete values. They found that the rating scale did not change the latent structure of responses and recommended the use of low-tech radio buttons.

5.3 Method

5.3.1 Participants

The study reported was embedded in a survey on health-related products and fielded from December 6 to December 31, 2001. Of the 2596 respondents who started the survey, 2340 (90.1%) reached the experimental section. Respondents' software was scanned for technological requirements (Microsoft Windows as operating system, Internet Explorer version 5.0 or higher as the actual browser, and Java installed). Of the 2340, 1926 participants (82.3%) met the required technological criteria and were randomly assigned to two independent experiments. In this paper we only refer to the experiment described above. Overall, we obtained a net sample of 779 cases (see Table 5.1) and a response rate within the experiment of 97%. Slightly more participants were male (50.8%) and the mean reported age was M = 43.0 (SD = 13.6) years.

Formal Education. All but seven respondents (99.1%) provided information on their education. We recoded the reported education according to the International Standard Classification of Education (ISCED, see United Nations Educational, Scientific and Cultural Organization, 1997) into two groups: Below ISCED 5A – ISCED 5A is considered as higher education and roughly starts with college degree, e.g. B.A. or B.S. – and ISCED 5A or higher.

Table 5.1

Assignment to Experimental Conditions

Rating scale	Alignment		Overall
	Horizontal	Vertical	
Slider scale	25.3% (*n* = 197)	24.9% (*n* = 194)	50.2% (*n* = 391)
Radio button scale	23.7% (*n* = 185)	26.1% (*n* = 203)	49.8% (*n* = 388)
Overall	49.0% (*n* = 382)	51.0% (*n* = 397)	100% (*N* = 779)

5.3.2 Procedure

The experimental treatment followed the presentation and evaluation of two abstract product concepts – *boost* ("increase both mental and physical energy for up to 3 hours") and *soothe* ("relax a person for up to 3 hours") – and appeared late in the questionnaire (mean duration of participation before the experiment *M* = 9.7 minutes, *SD* = 6.8). Both concepts were presented on a single Web page, counterbalanced for order. Ratings with radio button scales were mandatory. With slider scales the value of the middle category was submitted if respondents proceeded to the next page without changing the position of the handle.

5.4 Results

5.4.1 Break-off

Overall, 3.0% (*n* = 23) of the respondents quit participation during the experiment (see Table 5.2). Most of the break-off occurred in the slider scale condition, $\chi^2(1, N = 779) = 12.81$, $p < .001$, odds ratio = 6.92.

Break-Off and Formal Education. Within the group of respondents with a low formal education the probability for dropping out was higher with slider scales than with radio button scales, $\chi^2(1, N = 451) = 5.89$, $p = .018$, odds ratio = 5.45. When only looking at the group of respondents with a high formal education, Fisher's exact test did not detect any difference in break-off between slider scales and radio button scales, $\chi^2(1, N = 321) = 1.66$, $p = 1.000$. Spatial orientation did not statistically significantly impact break-off rate in any condition (see Table 5.2).

Table 5.2

Break-Off Within Each Condition, Overall, and per Formal Education

Rating scale	Dropout	No dropout
	Overall	
Slider scale – horizontal	5.6% (*n* = 11)[a]	94.4% (*n* = 186)
Slider scale – vertical	4.6% (*n* = 9)[b]	95.4% (*n* = 185)
Radio button scale – horizontal	1.6% (*n* = 3)[a]	98.4% (*n* = 182)
Radio button scale – vertical	0.0% (*n* = 0)[b]	100.0% (*n* = 203)
Total	3.0% (*n* = 23)	97.0% (*n* = 779)
	Education < ISCED 5A	
Slider scale – horizontal	4.9% (*n* = 5)	95.1% (*n* = 98)
Slider scale – vertical	4.3% (*n* = 5)[c]	95.7% (*n* = 112)
Radio button scale – horizontal	1.8% (*n* = 2)	98.2% (*n* = 107)
Radio button scale – vertical	0.0% (*n* = 0)[c]	100.0% (*n* = 122)
Total	2.7% (*n* = 11)	97.3% (*n* = 439)
	Education ≥ ISCED 5A	
Slider scale – horizontal	2.2% (*n* = 2)	97.8% (*n* = 88)
Slider scale – vertical	2.7% (*n* = 2)	97.3% (*n* = 73)
Radio button scale – horizontal	1.3% (*n* = 1)	98.7% (*n* = 75)
Radio button scale – vertical	0.0% (*n* = 0)	100.0% (*n* = 80)
Total	1.6% (*n* = 5)	98.4% (*n* = 5)

Note. Means with the same superscript differ statistically significantly according to chi-square tests/Fisher's exact test, only comparisons between same type of scale or same alignment were computed.
[a,c]$p < .05$. [b]$p \leq .001$.

Further analyses of response times and responses are based on the remaining 756 complete cases.

5.3.2 Task duration

Before analysis of response time (e.g. Heerwegh, 2003), we removed all outliers within each group according to Tukey's (1977) definition: Response times lower than the first quartile minus 1.5 interquartile ranges or higher than the third quartile plus 1.5 interquartile ranges were excluded. Overall, we disregarded 6.7% of response times for these analyses resulting in

704 cases. Neither type of rating scale nor spatial orientation had a statistically significant association with respondent exclusion, $\chi^2(3, N = 756) = 1.79$, $p = .616$.

Table 5.3
Mean Task Duration (SD) in Seconds by Condition, Outlier Excluded

	Alignment		
Rating scale	Horizontal	Vertical	Overall
Slider scale	33.1 (11.8)	37.8 (14.7)	35.4 (13.5)***
Radio button scale	14.2 (5.5)	17.3 (5.9)	15.8 (5.9)***
Overall	23.7 (13.2)***	26.9 (15.0)***	25.3 (14.2)***

***$p \leq .001$

Task duration (see Table 5.3) was considerably higher with slider scales than with radio buttons, $F(1, 703) = 638{,}23$, $p < .001$, $\eta^2 = .48$. Ratings on horizontal scales took statistically significant less time to complete than on vertical scales, but the effect size was small, $F(1, 703) = 8.81$, $p = .003$, $\eta^2 = .01$. No significant interaction between scale and spatial orientation occurred, $F(1, 703) = 1.02$, $p = .313$.

We analyzed the influence of spatial orientation separately within each condition because the bulky Java technology needed for the slider scales was likely to increase response times. In comparison to vertical alignment completion took less time with horizontal slider scales, $F(1, 341) = 10.41$, $p = .001$, $\eta^2 = .03$, as well as with horizontal radio button scales, $F(1, 362) = 26.70$, $p < .001$, $\eta^2 = .07$.

5.3.3 Content of responses

Mean Ratings. Table 5.4 shows the mean ratings of both concepts. Slider scales lead to higher ratings for measurement of both concepts. The difference was statistically significant for *boost*, $F(1, 754) = 7.48$, $p = .006$, $\eta^2 = .01$, but not for *soothe*, $F(1, 754) = 3.17$, $p = .076$. Spatial orientation did not have a statistically significant influence on ratings, neither for *boost*, $F(1, 754) < 1$, *ns*, nor for *soothe*, $F(1, 754) = 2.85$, $p = .092$.

Table 5.4

Mean Ratings (Standard Deviations) per Condition

Rating scale	Alignment		Overall
	Horizontal	Vertical	
	Concept *boost*		
Slider scale	4.2 (2.0)	4.2 (2.0)[a]	4.2 (2.0)[b]
Radio button scale	3.8 (1.9)	3.8 (1.9)[a]	3.8 (1.9)[b]
Total	4.0 (2.0)	4.0 (2.0)	4.0 (2.0)
	Concept *soothe*		
Slider scale	4.4 (2.0)	4.2 (2.1)	4.3 (2.0)
Radio button scale	4.2 (1.7)	4.0 (1.8)	4.1 (1.8)
Overall	4.3 (1.9)	4.1 (1.9)	4.2 (1.9)

Note. 1 = *very negative*, 7 = *very positive*. Means with the same superscript differ statistically significantly.
[a]$p < .05$. [b]$p < .01$

Mode. We found only small differences in mode values between slider scales and radio button scales. For measurement of the concept *boost* the mode is shifted one category to the positive end (right respectively bottom) of the scale with slider scales (category 6 instead of 5). Within each scale spatial orientation does not matter. Measurement of *soothe* leads to the same mode (category 6) for all rating scales but radio button scales' mode is shifted two categories to the negative (left respectively top).

Distribution of Values. Overall, rating scale has an impact on the distribution of values (see Figure 5.2) for *boost*, $\chi^2(6, N = 756) = 17.58$, $p = .007$, as well as for *soothe*, $\chi^2(6, N = 756) = 23.62$, $p = .001$. Alignment had neither an influence for the distribution of *boost*, $\chi^2(6, N = 756) = 3.71$, $p = .716$, nor for *soothe*, $\chi^2(6, N = 756) = 10.38$, $p = .110$. As standardized residuals suggest, especially differences in the middle response option make a major contribution to the overall difference between rating scales (see Figure 5.2).

Use of Middle Category. For concept *boost*, 10.5% selected the middle category with slider scales, whereas 15.8% gave a middle rating with radio button scales, $\chi^2(1, N = 756) = 4.68$, $p = .031$, odds ratio = 1.60. Results for measurement of the concept *soothe* are comparable: 11.6% chose the middle category with slider scales and 21.0% with radio button scales, $\chi^2(1, N = 756) = 12.30$, $p < .001$, odds ratio = 2.03. Spatial orientation did not matter

regarding the use of the middle category for both concepts, neither within nor between scales.

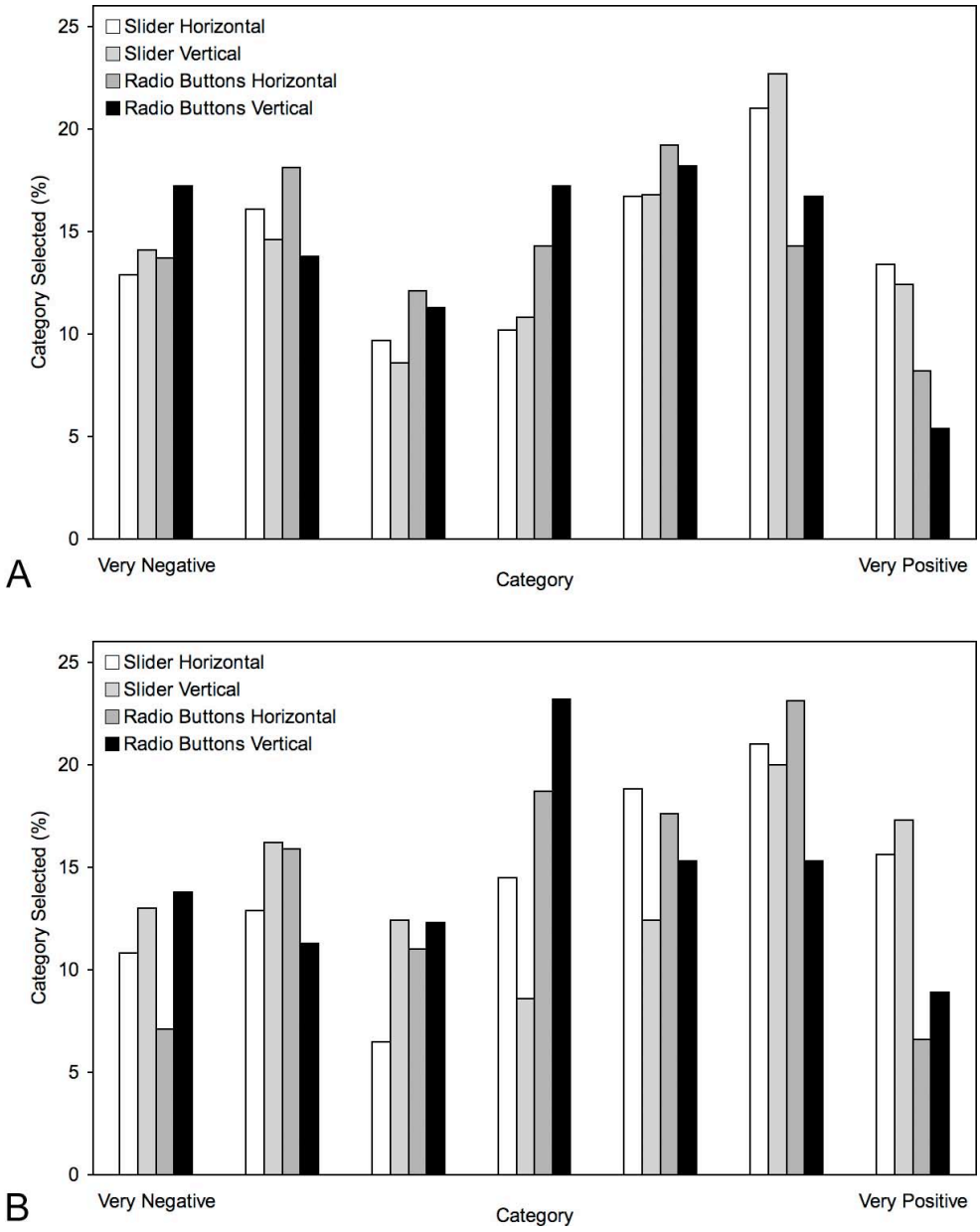


Figure 5.2. Distribution of values for measurement of concept *boost* (A) and concept *soothe* (B).

5.5 Summary

We compared Java-based slider scales and HTML radio buttons in horizontal versus vertical orientation in a Web-based questionnaire. Analyses were conducted both within the factor rating scales and within the factor spatial orientation. No interactions were observed at all.

5.5.1 Slider scales versus radio button scales

Break-Off and Education. The number of participants breaking off a study is a main indicator for task difficulty or technical problems and can be a major source of bias. Overall, slider scales produced about seven times more break-offs than radio button scales. We observed a strong relation between low formal education – a simplistic proxy for cognitive abilities – and break-off. This difference resonates well with the study by Buchanan and Reips (2001), where education and personality traits also interacted with technology to explain non-response behavior. The present results suggest that certain formats are more challenging in terms of previous knowledge needed or cognitive load.

Response Time. In the present study we take a longer duration not as an indicator for deeper cognitive processing but for problems on the stage of formatting the answer on the available rating scale (see Sudman, Bradburn, & Schwarz, 1996). Slider scales lead to a significant increase of response time with a very large effect size, regardless of respondent's level of formal education. This finding is in line with Couper et al. (2006) who found substantially higher response times with Java-based visual analogue scales (see Reips & Funke, 2008). It remains inconclusive if technical reasons (e.g. loading time of the Java applet) or if the different and more demanding handling of the slider account for the main share of variance in response times.

Distribution of Values. In comparison to measurement with radio button scales *fewer* respondents chose the middle category with slider scales for both evaluated concepts. This is especially remarkable as the center category was pre-selected with slider scales and if respondents proceeded with the questionnaire without changing the position of the slider this was treated as an answer of middle intensity. There are two contrasting post-hoc explanations for this. Either the respondents thought that the middle position was no valid choice and treated the scale not like a 7-point scale but like a 6-point scale. Or they felt forced to communicate an attitude tending toward one extreme. With the present data this issue remains inconclusive but it illustrates that making ratings on slider scales substantially differs from making ratings with radio button scales.

Central Tendency. We observed small upward shifts of mean ratings for measurement with slider scales in comparison to radio button scales. The difference, however, reached statistical significance only for one out of two items and the effect size was low, i.e. roughly half of a point on the 7-point scale. The consistently higher ratings with slider scales could be

the effect of respondents being reluctant to use the middle category and instead choosing a higher category.

5.5.2 Horizontal versus vertical alignment

The only observable effect of changing alignment of rating scales was that judgments on vertical scales took slightly longer on average than ratings on horizontal scales.

5.6 Conclusions

5.6.1 Keep scales simple

Possible anchoring effects, more demanding usage in terms of cognitive effort for scale administration, as well as problems with identifying non-attitudes and the initial position of the slider are serious negative scale characteristics of slider scales. These restrictions likely cannot be compensated for by a better technological implementation. Additionally, we found empirical evidence for higher response times and a higher break-off rate with respondents with a low formal education, which disqualifies slider scales for being used as rating scale. As these disadvantages outweigh the potential advantages by far the authors clearly advice against the use of Java-based slider scales in general. Less demanding techniques following a low-tech paradigm should be used whenever possible (see Reips, 2002b). HTML radio buttons should be used for to measure ordinal concepts and visual analogue scales for interval concepts (Reips & Funke, 2008).

5.6.2 Twist without risk

Overall, it seems that horizontal and vertical layout can be substituted mutually broading the possibilities for the layout of Web surveys without running the risk of negatively affecting the measurement process. This may be especially valuable for mobile Web surveys on handheld devices like cellular phones, where small screens require a careful use of space.

References

References

Aitken, R. C. B. (1969). Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine, 62,* 989–993.

Arnau, R. C., Thompson, R. L., & Cook, C. (2001) Do different response formats change the latent structure of responses? An empirical investigation using taxometric analysis. *Educational and Psychological Measurement, 61*, 23–44.

Averbuch, M., & Katzper, M. (2004). Assessment of visual analog versus categorical scale for measurement of osteoarthritis pain. *Journal of Clinical Pharmacology, 44,* 368–372.

Barak, A. (1999). Psychological applications on the Internet: A discipline on the threshold of a new millennium. *Applied and Preventive Psychology, 8,* 231–246.

Barak, A. (Ed.) (2008). *Psychological aspects of cyberspace: Theory, research, applications.* Cambridge: Cambridge University Press.

Batinic, B. (Ed.) (2000). *Internet für Psychologen* [Internet for psychologists] (2nd ed.). Göttingen: Hogrefe.

Batinic, B., Reips, U.-D., & Bosnjak, M. (Eds.) (2002). *Online social sciences*. Seattle: Hogrefe.

Bellamy, N., Campbell, J., & Syrotuik, J. (1999). Comparative study of self-rating pain scales in osteoarthritis patients. *Current Medical Research & Opinion, 15,* 113–119.

Best, S. J., & Krueger, B. S. (2004). *Internet data collection.* Thousand Oaks: Sage.

Birnbaum, M. H. (Ed.) (2000a). *Psychological experiments on the Internet*. San Diego: Academic Press.

Birnbaum, M. H. (2000b). SurveyWiz and FactorWiz: JavaScript Web pages that make HTML forms for research on the Internet. *Behavior Research Methods, Instruments, & Computers, 32,* 339–346.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55,* 803–832.

Birnbaum, M. H., & Reips, U.-D. (2005). Behavioral research and data collection via the Internet. In R. W. Proctor & K.-P. L. Vu (Eds.), *The handbook of human factors in Web design* (pp. 471–492). Mahwah, NJ: Erlbaum.

Bijur, P. E., Silver, W., & Gallagher, E. J. (2001). Reliability of the visual analog scale for measurement of acute pain. *Academic Emergency Medicine, 8,* 1153–1157.

Bosnjak, M. (2001). Participation in non-restricted Web surveys: A typology and explanatory model for item non-response. In U.-D. Reips, & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 193–208)*.* Lengerich: Pabst.

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions.: The definitive guide to questionnaire design.* San Francisco: Jossey-Bass.

Brunier, G., & Graydon, J. (1996). A comparison of two methods of measuring fatigue in patients on chronic haemodialysis: Visual analogue vs. Likert scale. *International Journal of Nursing Studies, 33,* 338–348.

Buchanan, T., & Reips, U.-D. (2001). Platform-dependent biases in Online Research: Do Mac users really think different? In K. J. Jonas, P. Breuer, B. Schauenburg & M. Boos (Eds.), *Perspectives on Internet Research: Concepts and Methods*. Retrieved from http://www.psych.uni-goettingen.de/congress/gor-2001/contrib/buchanan-tom

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68,* 57–80.

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2005). *Instructing Web and telephone respondent to report date answers in format desired by surveyor* (Technical Report 05-067). Retrieved from http://www.sesrc.wsu.edu/dillman/papers/Month%20Year%20Technical%20Report.pdf

Christian, L. M., Parsons, N. L., & Dillman, D. A. (2009). Designing scalar questions for Web surveys. *Sociological Methods & Research, 37,* 393–425.

Connes, B. (1972). The use of electronic desk computers in psychological experiments. *Journal of Structural Learning, 3,* 51–72.

Cork, R. C., Isaac, I., Elsharydah, A., Saleemi, S., Zavisca, F., & Lori A. (2004). A comparison of the verbal rating scale and the visual analog scale for pain assessment. *The Internet Journal of Anesthesiology, 8*(1). Retrieved from http://www.ispub.com/ostia/ index.php?xmlFilePath=journals/ija/vol8n1/vrs.xml

Couper, M. P. (2008). *Designing effective Web surveys.* Cambridge: Cambridge University Press.

Couper, M. P., Conrad, F. G., Tourangeau, R. (2007). Visual context effects in Web surveys. *Public Opinion Quarterly, 71,* 623–634.

Couper, M. P., Tourangeau, R., & Conrad, F. G. (2004). What they see is what we get: Response options for Web surveys. *Social Science Computer Review, 22,* 111–127.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review, 24,* 227–245.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65,* 230–253.

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.). (2008*). International handbook of survey methodology.* New York: Lawrence Erlbaum Associates.

Dillman, D. A., & Bowker, D. K. (2001). The Web questionnaire challenge to survey methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 159–178). Lengerich: Pabst.

Dillman, D. A., & Smyth, J. D. (2004). Design effects in the transition to Web-based surveys [Supplement]. *American Journal of Preventive Medicine, 5,* S90–S96.

Dillman D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken: Wiley.

Döring, N. (2003). *Sozialpsychologie des Internet* [Social psychology of the Internet] (2nd ed.). Göttingen: Hogrefe.

Downie, W. W., Leatham, P. A., Rhind, V. M., Wright, V., Branco, J. A., & Anderson, J. A. (1978). Studies with pain rating scales. *Annals of the Rheumatic Diseases, 37,* 378–381.

Eid, M., & Diener, E. (2006) (Eds.), *Handbook of multimethod measurement in psychology*. Washington: American Psychological Association.

Faas, T., & Schoen, H. (2006). Putting a questionnaire on the Web is not enough: A comparison of online and offline surveys conducted in the context of the German federal election 2002. *Journal of Official Statistics, 22,* 177–190.

Flynn, D., van Schaik, P., & van Wersch, A. (2004). A comparison of multi-item Likert and visual analogue scales for the assessment of transactionally defined coping function. *European Journal of Psychological Assessment, 20,* 49–58.

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks: Sage.

Fuchs, M. (2005, July). *Zur Messung von Häufigkeiten. Online-Befragung und Paper & Pencil-Befragung im Vergleich* [Measurement of frequencies: Online surveys compared to paper and pencil surveys]. Paper presented at the conference of the method section of the German Society of Sociology, Mannheim, Germany.

Fuchs, M. (2008). Mobile Web survey: A preliminary discussion of methodological implications. In F. Conrad & M. Schober (Eds.), *Envisioning the future of survey interviews* (pp. 77–94). New York: Wiley.

Fuchs, M., & Couper, M. P. (2001, September). *Length of input field and the responses provided in a self-administered survey: A comparison of paper and pencil and a Web survey.* Paper presented at the International Conference on Methodology and Statistics, Ljubljana, Slovenia.

Fuchs, M., & Funke, F. (2007). Multimedia Web surveys: Results from a field experiment on the use of audio and video clips in Web surveys. In M. Trotman et al. (Eds.), *The challenges of a changing world: Proceedings of the fifth international conference of the Association for Survey Computing* (pp. 63–80). Berkeley: ASC.

Funke, F. (2004). *Vergleich Visueller Analogskalen mit Kategorialskalen in Offline- und Onlinedesign* [Comparision of visual analogue scales and categorical rating scales in offline and online design]. Unpublished master's thesis. Justus Liebig University Gießen. Retrieved from http://frederikfunke.de/dateien/Funke_(2004)_Magisterarbeit.pdf

Funke, F., & Reips, U.-D. (2006, March). *Visual analogue scales in online surveys: Non-linear data categorization by transformation with reduced extremes*. Poster presented at the General Online Research conference, Bielefeld, Germany.

Funke, F., & Reips, U.-D. (2007a, March). *Dynamische Formulare: Onlinebefragungen 2.0* [Dynamic forms: Online surveys 2.0]. Paper presented at the General Online Research conference, Leipzig, Germany.

Funke, F., & Reips, U.-D. (2007b). Datenerhebung im Netz: Messmethoden und Skalen [Data collection on the Web: Measurement and rating scales]. In M. Welker & O. Wenzel (Eds.), *Online-Forschung 2007: Grundlagen und Fallstudien* (pp. 52–76). Köln: Halem.

Funke, F., & Reips, U.-D. (2008, March). *Differences and similarities between visual analogue scales, slider scales and categorical scales in Web surveys.* Poster presented at the General Online Research conference, Hamburg (Germany).

Funke, F., & Reips, U.-D. (2010a). *Formatting error with visual analogue scales and ordinal rating scales.* Manuscript submitted for publication.

Funke, F., & Reips, U.-D. (2010b). *Semantic differentials made from visual analogue scales: High data quality with elaborate and spontaneous judgments.* Manuscript submitted for publication.

Funke, F., Reips, U.-D., & Thomas, R. K. (in press). Sliders for the smart: Type of rating scale on the Web interacts with educational level. *Social Science Computer Review.*

Galesic, M., Tourangeau, R., Couper, M., & Conrad, F. (2008). Eye-tracking data: New insights on response order effects and other signs of cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72,* 892–913.

Gerich, J. (2007). Visual analogue scales for mode-independent measurement in self- administered questionnaires. *Behavior Research Methods, 39,* 985–992.

Gnambs, T. (2008). *Graphische Analogskalen in Internet-basierten Fragebögen* [Graphic analogue scales in Internet-based questionnaires]. München: Akademischer Verlag.

Göritz, A. S. (2006). Incentives in Web studies: Methodological issues and a review. *International Journal of Internet Science, 1*, 58–70.

Göritz, A. S., & Birnbaum, M. (2005). Generic HTML Form Processor: A versatile PHP script to save Web-collected data into a MySQL database. *Behavior Research Methods, 37,* 703–710.

Gosling, S. & Johnson, J. (Eds.) (2010). *Advanced Internet methods in the behavioral sciences*. Washington: American Psychological Association.

Graham, N. E. (1956). The speed and accuracy of reading horizontal, vertical, and circular scales. *Journal of Applied Psychology, 40*(4), 228–232.

Gräf, L. (1999). Optimierung von WWW-Umfragen: Das Online-Pretest-Studio [Optimizing WWW surveys: The Online-Pretest-Studio]. In B. Batinic, A. Werner, L. Gräf, & W. Bandilla (Eds.), *Online Research: Methoden, Anwendungen und Ergebnisse* (pp. 159–177). Göttingen: Hogrefe.

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodolody* (2nd ed.). Hoboken: Wiley.

Hartig, J., Jude, N., & Rauch, W. (2003). Entwicklung und Erprobung eines deutschen Big-Five-Fragebogens auf Basis des International Personality Item Pools (IPIP40) [Development and test of a German Big Five questionnaire based on the International Person-

ality Item Pool (IPIP40)]. Frankfurt/M.: *Arbeiten aus dem Institut der J. W. Goethe-Universität,* 2003/1.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin, 18,* 98–99.

Healey, B. (2007). Drop downs and scroll mice: The effect of response option format and input mechanism employed on data quality in Web surveys. *Social Science Computer Review, 25,* 111–128.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client side paradata from a Web survey. *Social Science Computer Review, 21,* 360–373.

Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in Web surveys. *Social Science Computer Review, 20*, 471–484.

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring visual analogue scales. *British Journal of Mathematical and Statistical Psychology, 61,* 401–413.

Hoggatt, A. C. (1977). On the uses of computers for experimental control and data acquisition. *American Behavioral Scientist, 20,* 347–365.

Honing, H., & Reips, U.-D. (2008). Web-based versus lab-based studies: A response to Kendall (2008). *Empirical Musicology Review, 3,* 73–77.

Jäger, R. (2004). Konstruktion einer Ratingskala mit Smilies als symbolische Marken [Construction of a rating scale with smileys as symbolic markers]. *Diagnostica, 50,* 31–38.

Janetzko, D., Hildebrandt, M., & Meyer, H. A. (2003). *Das Experimentalpsychologische Praktikum im Labor und WWW* [Experimental psychology studies in the lab and on the Web]. Götingen: Hogrefe.

Jensen, M. P., Turner, J. A., & Romano, J. M. (1994). What is the maximum number of levels needed in pain intensity measurement? *Pain, 58,* 387–392.

Joinson, A. N., McKenna, K., Postmes, T., & Reips, U.-D. (Eds.). (2007). *The Oxford handbook of Internet psychology*. Oxford: Oxford University Press.

Joinson, A. N., Woodley, A., & Reips, U.-D. (2007). Personalization, authentication and self-disclosure in self-administered Internet surveys. *Computers in Human Behavior, 23,* 275–285.

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychology Experiments on the Internet* (pp. 35–60). San Diego: Academic Press.

Kreindler, D., Levitt, A., Woolridge, N., & Lumsden, C. J. (2003). Portable mood mapping: The validity and reliability of analog scale displays for mood assessment via hand-held computer. *Psychiatry Research, 120,* 165–177.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537–567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51,* 201–219.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141–164). New York: Wiley.

Likert, R. (1932). *A technique for the measurement of attitudes.* New York: Columbia University.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4,* 73–79.

Lyberg, L., Biemer, P., Collins, M., deLeeuw, E., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). (1997). *Survey measurement and process quality.* New York: Wiley.

Mangan, M., & Reips, U.-D. (2007). Sleep, sex, and the Web: Surveying the difficult-to-reach clinical population suffering from sexsomnia. *Behavior Research Methods, 39,* 233–236.

Marcell, M. M., & Falls, A. L. (2001). Online data collection with special populations over the World Wide Web. *Down Syndrome Research and Practice, 7,* 106–123.

Marsh-Richard, D., Shannon, E., Mathias, C., Venditti, N., & Dougherty, D. (2009). Adaptive visual analog scales (AVAS): A modifiable software program for the creation, administration, and scoring of visual analog scales. *Behavior Research Methods, 41,* 99–106.

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). Psyc-Exps: An online psychological laboratory. In M.H. Birnbaum (Ed.), *Psychology Experiments on the Internet* (pp. 219–233). San Diego: Academic Press.

McReynold, P., & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences, 8,* 281–283.

Mühlenfeld, H.-U. (2004). *Der Mensch in der Online-Kommunikation: Zum Einfluss webbasierter, audiovisueller Fernkommunikation auf das Verhalten von Befragten* [Human aspects of online communication: The influence of Web-based audio-visual communication on respondent behavior]. Wiesbaden: DUV.

Musch, J., & Reips, U.-D. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–88). San Diego: Academic Press.

Myles, P. S., Troedel, S., Boquest, M., & Reeves, M. (1999). The pain visual analog scale: Is it linear or nonlinear? *Anesthesia & Analgesia, 89,* 1517–1520.

Myles P. S., & Urquhart N. (2005). The linearity of the visual analogue scale in patients with severe acute pain. *Anaesthesia and Intensive Care, 33,* 54–58.

Neubarth, W. (2006, March). *Ranking vs. rating in an online environment*. Paper presented at the General Online Research conference, Bielefeld, Germany.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49,* 197–237.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbama, IL: University of Illinois Press.

Peytchev, A., & Hill, C. A. (2010). Experiments in mobile Web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review, 28*.

Proctor, R. W., & Vu, K.-P. L. (Eds.) (2005). *The handbook of human factors in Web design*. Mahwah, NJ: Erlbaum.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–118). San Diego, CA: Academic Press.

Reips, U.-D. (2001). The Web Experimental Psychology Lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, & Computers, 33,* 201–211.

Reips, U.-D. (2002a). Context effects in Web surveys. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 69–80). Seattle: Hogrefe & Huber.

Reips, U.-D. (2002b). Internet-based psychological experimenting: Five dos and five don'ts. *Social Science Computer Review, 20,* 241–249.

Reips, U.-D. (2002c). Standards for Internet-based experimenting. *Experimental Psychology, 49*, 243–256.

Reips, U.-D. (2003). Psychologische Forschung zum und im Internet [Psychological research on and in the Internet]. *Psychologie in Österreich, 22*(1), 19–25.

Reips, U.-D. (2006a). Computer-vermittelte Kommunikation [Computer-mediated communication]. In H. W. Bierhoff & D. Frey (Eds.), *Handbuch der Sozialpsychologie und Kommunikationspsychologie* (pp. 555–564). Göttingen: Hogrefe.

Reips, U.-D. (2006b). Internet-basierte Methoden [Internet-based methods]. In F. Petermann & M. Eid (Eds.), *Handbuch der Psychologischen Diagnostik* (pp. 218–225). Göttingen: Hogrefe.

Reips, U.-D. (2006c). Web-based methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 73–85). Washington: American Psychological Association.

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of Internet psychology* (pp. 373–390). Oxford: Oxford University Press.

Reips, U.-D. (2010). Design and formatting in Internet-based research. In S. Gosling & J. Johnson, *Advanced Internet methods in the behavioral sciences* (pp. 29–43). Washington, DC: American Psychological Association.

Reips, U.-D., & Birnbaum, M. H. (in press). Behavioral research and data collection via the Internet. In R. W. Proctor & K.-P. L. Vu (Eds.), *The handbook of human factors in Web design* (2nd ed.). Mahwah, NJ: Erlbaum.

Reips, U.-D., & Bosnjak, M. (Eds.). (2001). *Dimensions of Internet science*. Lengerich: Pabst.

Reips, U.-D., & Funke, F. (2008). Interval level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods, 40*, 699–704.

Reips, U.-D., & Krantz, J. (2010). Conducting true experiments on the Web. In S. Gosling & J. Johnson, *Advanced Internet methods in the behavioral sciences* (pp. 193–216). Washington: American Psychological Association.

Reips, U.-D., & Lengler, R. (2005). The Web Experiment List: A Web service for the recruitment of participants and archiving of Internet-based experiments. *Behavior Research Methods, 37,* 287–292.

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, Instruments, & Computers, 34,* 234–240.

Reips, U.-D., & Stieger, S. (2004). Scientific LogAnalyzer: A Web-based tool for analyses of server log files in psychological research. *Behavior Research Methods, Instruments, & Computers, 36,* 304–311.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84,* 754–775.

Sassenberg, K., Postmes, T., Boos, M., & Reips, U.-D. (2003). Studying the Internet: A challenge for modern psychology. *Swiss Journal of Psychology, 62,* 75–77.

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers, 29,* 274–279.

Schmidt, W. C. (2007). Technical considerations when implementing online research. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of Internet psychology* (pp. 459–470). Oxford: University Press.

Schonlau, M., Asch, B. J., & Du, C. (2003). Web surveys as part of a mixed-mode strategy for populations that cannot be contacted by e-mail. *Social Science Computer Review, 21,* 218–222.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105.

Schwarz, N., Hippler, H.-J, Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49,* 388–395.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55,* 570–582.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22,* 127–160.

Schwarz, S., & Reips, U.-D. (2001). CGI versus JavaScript: A Web experiment on the reversed hindsight bias. In U.-D. Reips, & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 75–90). Lengerich: Pabst.

Scott, J. & Huskisson, E. C. (1979). Vertical or horizontal visual analogue scales. *Annals of the Rheumatic Diseases, 38,* 560.

Seymour, R. A., Simpson, J. M., Charlton, J. E., & Phillips, M. E. (1985). An evaluation of length and end-phrase of visual analogue scales in dental pain. *Pain, 21,* 177–185.

Skitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology, 57,* 529–555.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2004). How visual grouping influences answers to Internet surveys. Retrieved from http://www.sesrc.wsu.edu/ dillman/papers/(6_1)%20Grouping%20paper%20for%20binding.pdf

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006a). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly, 70,* 66–77.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006b). Effects of using visual design principles to group response options in Web surveys. *International Journal of Internet Science, 1,* 6–16.

Stieger, S., & Reips, U.-D. (in press). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*.

Stieger, S., Reips, U.-D., & Voracek, M. (2007). Forced response in online surveys: Bias from reactance and an increase in sex-specific dropout. *Journal of the American Society for Information Science and Technology, 58,* 1653–1660.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretative heuristics for visual features of survey questions. *Public Opinion Quarterly, 68,* 368–393.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007). Color, labels, and interpretative heuristics for response scales. *Public Opinion Quarterly, 71*, 91–112.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tiplady, B., Jackson, S. H. D., Maskrey, V. M., & Swift, C. G. (1998). Validity and sensitivity of visual analogue scales in young and older healthy subjects. *Age and Aging, 27,* 63–66.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpreteative heuristics for visual features of survey questions. *Public Opinion Quarterly, 68,* 368–393.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Turpin, J., Rose, R., & Larsen, B. (2003). An adaptive, transportable Web-based data acquisition platform for Clinical and Survey-Based Research. *Journal of the American Osteopathic Association, 103,* 182–186.

United Nations Educational, Scientific and Cultural Organization (1997). *International standard classification of education – 1997 version.* Retrieved from http://www.uis.unesco.org/TEMPLATE/pdf/isced/ISCED_A.pdf

van Schaik, P., & Ling, J. (2007). Design parameters of rating scales. *ACM Transactions on Computer-Human Interaction, 14*(1), 4.

Welker, M., & Wenzel, O. (Eds.), (2007). *Online-Forschung 2007: Grundlagen und Fallstudien* [Online reseach 2007: Basics and case studies]. Köln: Halem.

Appendix A

Conference Presentations

Appendix A: Conference Presentations

In the context of this dissertation the following 22 presentations have been given at 19 conferences and workshops between 2005 and 2010. Slides are available on the author's Web site (http://frederikfunke.de) or on request.

Fuchs, M., & Funke, F. (2007, September). *Multimedia Web surveys: Results from a field experiment on the use of audio and video clips in Web surveys.* Paper presented at the 5th international conference on Survey and Statistical Computing of the Association for Survey Computing (ASC), Southampton (GB).

Fuchs, M., & Funke, F. (2007, September). *Using audio and video clips in Web surveys: Feasibility and impact on data quality.* Paper presented at the 60th annual conference of the World Association for Public Opinion Research (WAPOR), Berlin.

Fuchs, M., & Funke, F. (2007, September). *Video Web surveys: Results from field-experimental comparisons with text-based Web surveys.* Paper presented at the 2007 Workshop on Internet Survey Methodology, Lillehammer (Norway).

Funke, F. (2005, March). *Visual analogue scales in online surveys.* Paper presented at the 7th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Zurich (Switzerland).

Funke, F. (2007, September). *Data from visual analogue scales: Quality and comparison to categorical scales.* Paper presented at the 2007 Workshop on Internet Survey Methodology, Lillehammer (Norway).

Funke, F., & Reips, U.-D. (2005, July). *Stichprobenverzerrung durch browserbedingten Dropout* [Sample bias by browser caused dropout]. Paper presented at the conference of the method section of the German Society for Sociology (DGS), Mannheim (Germany).

Funke, F., & Reips, U.-D. (2007, March). *Dynamic forms: Online surveys 2.0.* Paper presented at the 9th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Leipzig (Germany).

Funke, F. & Reips, U.-D. (2007, June). *Improving data quality in Web surveys with visual analogue scales.* Paper presented at the conference of the European Survey Research Association (ESRA), Prague (CZ).

Funke, F., & Reips, U.-D. (2007, November). *Evaluating the influence of visual analogue scales on response behavior.* Paper presented at the 37th annual conference of the Society for Computers in Psychology (SCiP), Long Beach, CA (USA).

Funke, F., & Reips, U.-D. (2008, March). *Visual analogue scales versus categorical scales: Respondent burden, cognitive depth, and data quality.* Paper presented on the 10th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Hamburg.

Funke, F., & Reips, U.-D. (2008, July). *Assessment with visual analogue scales on the Internet.* Paper presented at the 29th International Congress of Psychology ICP 2008, Berlin.

Funke, F., & Reips, U.-D. (2008, September). *Assessing semantic differentials with visual analogue scales in Web surveys.* Paper presented at the 7th international conference of the RC-33 (Research Committee 33 "Logic and Methodology") of the International Sociological Association (ISA), Naples (Italy).

Funke, F., & Reips, U.-D. (2009, April). *Formatting error with visual analogue scales in Web surveys.* Paper presented at the 11th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Vienna (Austria).

Funke, F., & Reips, U.-D. (2009, April). *Results from 6 independent Web experiments comparing visual analogue scales with categorical scales.* Paper presented at the 11th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Vienna (Austria).

Funke, F., & Reips, U.-D. (2009, July). *Twisting rating scales: Horizontal versus vertical visual analogue scales versus categorical scales in Web-based research.* Paper presented at the 3rd conference of the European Survey Association (ESRA), Warsaw (Poland).

Funke, F., & Reips, U.-D. (2009, September). *Internet survey methods: Evolutions and revolutions.* Paper presented at the 1st International Workshop on Internet Survey, Deajeon (South Korea).

Funke, F., & Reips, U.-D. (2009, September). *Yes, VASs can! Visual analogue scales for Web-based research.* Paper presented at the 2009 ISM Workshop on Internet Survey Methodology, Bergamo (Italy).

Funke, F., & Reips, U.-D. (2010, May). *Formatting error with ordinal rating scales and visual analogue scales.* Paper presented at the 12th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Pforzheim.

Funke, F., & Reips, U.-D. (2010, July). *Detecting small effects with visual analogue scales.* Paper presented at the XVII ISA World Congress of Sociology, Gothenburg (Sweden).

Funke, F., Reips, U.-D., & Thomas, R. K. (2008, September). *Visual analogue scales in cross cultural Web surveys.* Paper presented at the 7th international conference of the RC-33 (Research Comittee 33 "Logic and Methodology") of the International Sociological Association (ISA), Naples (Italy).

Funke, F., Reips, U.-D., & Thomas, R. K. (2009, July). *Increasing confidence in survey estimates with visual analogue scales.* Paper presented at the 3rd conference of the European Survey Association (ESRA), Warsaw (Poland).

Thomas, R. K., Terhanian, G., & Funke, F. (2009, April). *Response formats in cross-cultural comparisons in Web-based surveys.* Paper presented at the 11th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Vienna (Austria).

Appendix B

Poster Presentations

Appendix B: Poster Presentations

The following posters are also available on the author's Web site (http://frederikfunke.de) or on request.

Funke, F., & Reips, U.-D. (2006a, March). *Visual analogue scales: Non-linear data categorization by transformation with reduced extremes.* Poster session presented at the 8th annual General Online Research (GOR) conference of the German Society for Online Research (D.G.O.F.), Bielefeld (Germany).

F. Funke[a] & U.-D. Reips[b] (University of Zurich)

Visual Analogue Scales in Online Surveys: Non-Linear Data Categorization by Transformation with Reduced Extremes

G.O.R.06

Universität Zürich

Introduction

Visual Analogue Scales (VAS) are graphic rating scales. They were first described by Hayes and Patterson in 1921. In most cases, a VAS is a simple horizontal line with verbal anchors on each end. Respondends convey their attitude or level of accordance by marking the point on the line they think is most appropriate. These scales have proven a highly reliable and valid instrument in surveys (Flynn et al., 2004).

Online surveys lend themseves to the use of VAS (see Figure 1). In contrast to paper and pencil interviews – where reading out data takes up a lot of time and is prone to errors – the readout in online surveys occurs automatically. The only client-side requirement is that the technology used to create the scales (e.g. JavaScript, Flash or Java) has to be enabled in the user's web browser.

Figure 1: Visual Analogue Scale in an online survey

From a statistical point of view, the main advantage is that VAS generate data on interval scale level and therefore meet important requirements for the applicability of parametric procedures (Nyrén et al., 1987). To understand the influence of VAS on the respondent's way of answerring questions we conducted two experiments.

Experiment 1 - Comparing VAS with Categorical Scales

In our first web experiment 667 participants were randomized to three different conditions to rate 16 items on power motivation (Schmidt & Frieze, 1997). The only difference between the experimental conditions consisted in the applied rating scales: Either a 4-point categorical scale, a 8-point categorical scale or a VAS.

Figure 2: Linear transformation of VAS to 4-point scale

To be able to compare the frequencies of categorical scales with data from VAS, VAS values need to be categorized. The commonly used way of categorizing VAS is linear transformation. Here, equal intervals form one category. Figure 2 shows linear transformation from VAS to a 4-point scale.

When comparing the frequencies between categorized VAS and the original categorical scale, we found typical systematic differences in the distribution, especially concerning the extreme categories (Figure 3).

Figure 3: Percent categorical scale and VAS after linear transformation

To gain greater accordance, we categorized the VAS in a different way: „Transformation with reduced extremes" (Figure 4).

Figure 4: Transformation with reduced extremes

All in all, transformation with reduced extremes leads to high correspondence with categorical scales (Figure 5) and is superior to linear transformation.

Figure 5: Percent categorical scale and VAS after transformation with reduced extremes

In turn this means that the exteme categories of categorical scales represent a smaller intervall than the other categories; they are less frequently used by the respondents. The centers of categories and categorical scales have different distances, they are not perceived as equidistant.

Experiment 2 - Modifying Categorical Scales

To compensate for the findings of the first experiment that extreme categories represent smaller intervalls of intensity, we modified the categorical scales in the second experiment. The space between the extreme categories and the adjoining ones was scaled down (Figure 6). We expected these changes to influence participants' preference for extreme categories.

Figure 6: Categorical scales with different spacings and VAS used in experiment two

The participants (n=185) were randomized to one of these seven different scales to rate 58 items on 29 pages. The frequencies of scales with the same number of categories were compared to a VAS that was categorized linearly (Figure 7) to the respective number of categories.

Figure 7: Percent of categorical scales with equal spacings, categorical scales with reduced extremes and linearly transformed VAS

The manipulation of the extreme spacings had a clear influence on the observed frequencies. This effect is independent of the number of categories and even the slight divergence between the two variations of the 9-point scale made a difference. When using categorical scales with reduced extremes, the frequencies of extreme categories decreased.

Conclusion

A general conclusion that can be drawn is that it makes a difference for the distribution of frequencies whether items are rated by categorical scales or by VAS. In both experiments, the categorization of VAS according to the model of reduced extremes is superior to linear transformation. This means: if one wants to compare frequencies of data obtained from VAS with categorical data, linear transformation is inappropriate and categorization with reduced extremes should be applied.

As modification of spacings between categories has shown to produce robust effects that result in changes in the distribution of frequencies, upcoming research will focus on experiments with spacings between categories. In a further experimental design, the influence of categorical scales with extended extremes will be examined (Figure 8).

Figure 8: Categorical scales with extended extremes and VAS

References:

Hayes, M. H. S. & D. G. Patterson (1921): Experimental development of the graphic rating method: *Psychological Bulletin, 18*, 98-99

Flynn, D., P. van Schaik & A. van Wersh (2004). A Comparison of Multi-item Likert and Visual Analogue Scales for the Assessment of Transactional Web Site Quality Function. *European Journal of Psychological Assessment, 20*, 49-58

Nyrén, O. et al. (1987): Self-rating of pain in non-ulcer dyspepsia. *Journal of Clinical Gastroenterology, 9*, 408-414

Schmidt L. C. & I. H. Frieze (1997): A mediational model of power, affiliation, and achievement motives, and power involvement. *Journal of Business and Psychology, 4*, 425-446

[a] email@frederikfunke.de
[b] u.reips@psychologie.unizh.ch

Reips, U.-D., & Funke, F. (2006b, November). *Visual analogue scales for Internet-based research: VAS Generator*. Poster session presented at the 36th annual conference of the Society for Computers in Psychology (SCiP), Houston, TX (USA).

Funke, F., & Reips, U.-D. (2007, September). *New types of measurement in Internet-based research.* Poster session presented at the 10th congress of the Swiss Society for Psychology, Zurich (Switzerland).

New types of measurement in Internet-based research

Frederik Funke & Ulf-Dietrich Reips (University of Zurich)

University of Zurich

As examples for the many different new developments in the field of Internet-based research, we describe two types of measurement in self-administered questionnaires, visual analogue scales and dynamic lists. Whereas the former are well-known from paper-based questionnaires, the latter are a truly genuine Web methodology. What applies to both is that only with the transition to Web the surveying of large samples is feasible.

Type 1: visual analog scales – measurement on the level of an interval scale

Visual analog scales (VAS) are rating scales, used in self-administered questionnaires in paper and pencil studies as well as in computerized laboratory settings. Respondents can adjust the extent of the attitude, judgment, or impression being measured by clicking on the horizontal line between the verbal descriptors of the extremes. VAS are considered a reliable instrument for valid measurements. Automation in combination with a Web interface allows the fast and precise readout of data (see Figure 1). A free Web service (maintained by the authors) to create, test and download VAS for one's own studies is available at http://www.vasgenerator.net.

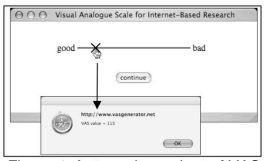


Figure 1. Automatic readout of VAS

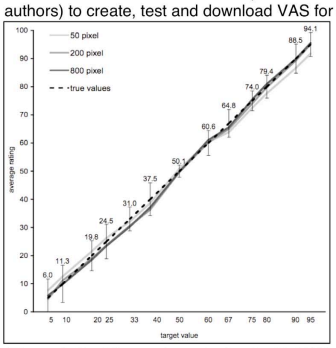


Figure 2. Equidistant data obtained with VAS

Empirical test of interval level measurement

To test if respondents use the scale in the intended way, so that data are on the desired level of an interval scale, we conducted a Web experiment with an Internet sample of 255 students. There were three different length conditions, one regular (200 pixels) and two extreme (50 pixels and 800 pixels), see lines below. Participants were instructed to repeatedly identify 13 target values on a VAS. As crucial indicators for interval level we checked for difference from true value and for equidistance, i.e. even numerical difference between values correspond with even distances on the scale.

Findings

Figure 2 shows the relation between target values and average ratings. On average the difference to a linear relationship was at 3.2 percentage points, ranging from 2.8 for the medium VAS to 3.9 for the shortest VAS. Because equal numerical intervals corresponded to roughly equal segments on the VAS and only minor aberration from true values was found across the scale even for extremely short and extremely long VAS, there is strong evidence that data collected with VAS are equidistant and on the level of an interval scale. Therefore, a wide range of statistical procedures can safely be applied when analyzing data measured with VAS (see Reips & Funke, in press).

– 50 pixels –
200 pixels
800 pixels

Type 2: dynamic lists – filtering closed-ended questions on the fly

Dynamic lists (Figure 3) assist respondents with finding the appropriate answer in closed-ended questions with many possible values. With dynamic elements, the respondent is guided through a hierarchical answering process, on a single Web page. The answering process is broken down into multiple steps. On load of the Web page only a very general choice is requested from the respondent. Immediately afterwards, further possible choices on the next, more specific level appear. Finally, after the second choice the final, very specific choices appear.

Web experiment

In an experimental design with three conditions we compared a dynamic list with conventional multi-page filtering (each level is presented on a separate Web page; Figure 4) and no filtering at all (i.e. all possible values were presented on a single page; Figure 5). In each condition respondents (n = 252) had to choose one from 48 possible values.

Findings

The most general finding is that dynamic lists work: we did not observe negative effects in the form of an increase of item nonresponse or dropout.

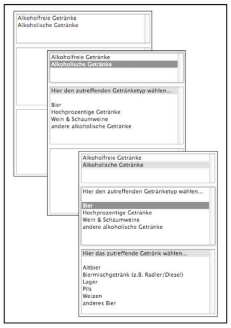


Figure 3. Condition 1: dynamic list on a single Web page

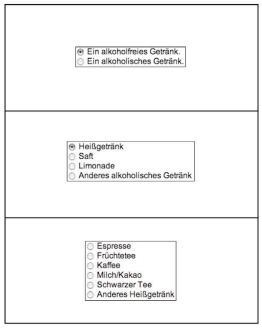


Figure 4. Condition 2: multipage filtering on 3 separate Web pages

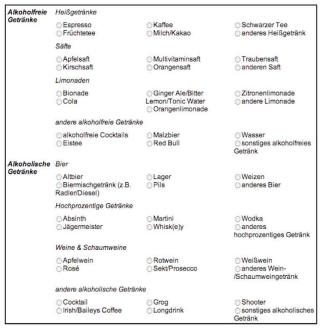


Figure 5. Condition 3: no filtering on a single Web page

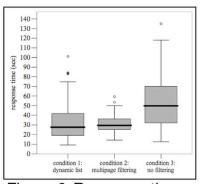


Figure 6. Response times

Furthermore, dynamic lists combine the advantages of knowing all choices with the speed of multipage filtering. Regarding response time (see Figure 6) we found the time needed to answer the question to be significantly higher than conditions 1 and 2 ($p < .001$) when all choices were presented on a single Web page (condition 3: $M = 52$ sec, $SD = 25$ sec). The variance was also larger, $W(1, 171) = 5.072$, $p < .05$, possibly indicating more difficulties with this format. The difference between dynamic list (condition 1: $M = 33$ sec, $SD = 20$ sec) and multipage filtering (condition 2: $M = 31$ sec, $SD = 9$ sec) was not statistically significant. Regarding the variety of answers we found that answers from the condition without filtering – where a deliberate choice could take place, as all possible answers were evident – resembled more the answers obtained with dynamic lists than with multipage filtering (see Funke & Reips, 2007). Thus, dynamic lists perform best regarding both, data quality and respondents burden (operationalized through response time).

Conclusion: Internet-based research expands methodological possibilities

As our examples show, new types of online measurement have the potential to improve data collection in speed, accuracy, data quality, and respondents burden. However, as previous research (e.g. Dillman, 2007) has shown, one has to be cautious when implementing new methods, as even minor changes can have a significant influence on data.

References

Dillman, D. A., & Smyth, J. D. (2007). Design effects in the transition to Web-based surveys. *American Journal of Preventive Medicine, 32,* 90-96.

Funke, F., & Reips, U.-D. (2007). Datenerhebung im Netz: Messmethoden und Skalen [Data collection in the Web: Measuring devices and scales]. In M. Welker & O. Wenzel (Eds.): *Onlineforschung 2007: Grundlagen und Fallstudien* (pp. 51-75). Cologne: Halem.

Reips, U.-D., & Funke, F. (in press). Interval level measurement with visual analog scales in Internet based research: VAS Generator. *Behavior Research Methods.*

Funke, F., & Reips, U.-D. (2008, March). *Differences and similarities between visual analogue scales, slider scales and categorical scales in Web surveys.* Poster session presented on the 10th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Hamburg (Germany).

Frederik Funke[1] & Ulf-Dietrich Reips[2]
[1]University of Tübingen, Germany; http://www.frederikfunke.de; email@frederikfunke.de; [2]University of Zurich, Switzerland; u.reips@psychologie.uzh.ch

GOR 08 Eberhard Karls Universität Tübingen
University of Zurich

Difference and Correspondences Between Visual Analogue Scales, Slider Scales and Radio Button Scales in Web Surveys

Research interest

We conducted a Web experiment to compare data and paradata from three scales (see Figure 1) for answering closed-ended questions in self-administered Web questionnaires: visual analogue scales (VAS), slider scales (SLS) and radio button scales (RBS). VAS are nearly continuous measurement instruments, each pixel is clickable and results in a raw value. In contrast, RBS only provide a limited number of categories. As we know from previous research, data from VAS reach the desired level of an interval scale (Reips & Funke, in press) and there is a systematic difference between the scales (especially concerning the extremes; see Funke & Reips, 2007). SLS are an in-between answer format. They are visually similar to VAS, but functionally more similar to the RBS. Do the scales have a different influence on data collection and data quality? Does usage follow function or appearance?

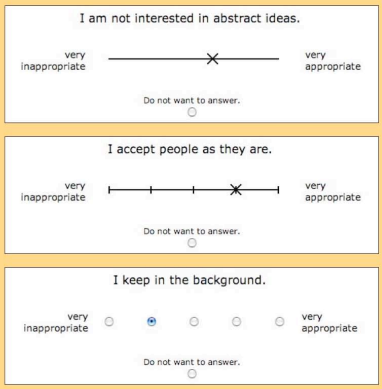


Figure 1. Scales compared in this study: VAS, SLS and RBS (from top to bottom)

Design

Respondents of a 40 item personality inventory (validated for Internet usage by Hartig, Jude, & Rauch, 2003) were randomly assigned to either a VAS with 250 possible values, an SLS with 5 discrete values, or a 5 point RBS. Directly after load neither scale provided an initial marker (with VAS and SLS the marker appeared after the first click on the scale). The VAS used in this survey could only be clicked (just as the RBS) whereas the SLS' marker could be clicked or slid but its final position was limited to the indicated discrete values.

Empirical data

As shown in Table 1, completion rate and missing data were not statistically significant different for the three answer formats, but there was a tendency that VAS performed better on these indicators than SLS and RBS (in contrast to findings by Couper, Tourangeau, Conrad, & Singer, 2006). As expected (see Couper et al., 2006) response time was higher with VAS than with RBS.
Means were different, $F(2, 279) = 5.67$, $p = .004$, with VAS measuring lower mean scores. We were able to determine test-retest reliability with a smaller sample and found a statistically significant difference between the scales, $V(2, 3280) = 35.7$, $p < .001$, with VAS producing the highest reliability scores.
Finally, we compared the distributions of values. Therefore we transformed data from VAS into 5 categories consisting of equal data intervals (see Figure 2). Our finding: SLS and RBS are used in a similar way, the distribution of VAS' frequencies differs substantially.

Table 1. Indicators for data quality

Indicator	VAS	SLS	RBS	Total
	Paradata			
Completion rate[a]	97%	95%	94%	96% n.s.
n (net sample)[b]	107	87	88	282
Mean missing data rate[b]	0.6%	0.9%	0.9%	0.8% n.s.
Mean response time per item in seconds (*SD*)[b, c]	7.3 (2.0)	6.9 (1.9)	6.6 (1.5)	6.9 (1.8)*
	Central tendency[c]			
M (*SD*) for all 40 items	2.84 (0.17)	2.92 (0.16)	2.92 (0.20)	2.89 (0.18)**
	Test-retest reliability[e]			
n (net sample)	32	23	27	82
Mean reliability for all 40 items	.88	.82	.83	.84***

[a]Serious respondents only. [b]Serious and complete respondents only. [c]Adjustment: unreasonably high response times (> 60 sec) and outlier (i.e. not within *M* +/-2.5 interquartile ranges) were removed. [d]Scales running from1 to 5. [e]Two waves; complete and serious respondents only.
*p < .05. **p < .01. ***p < .001.

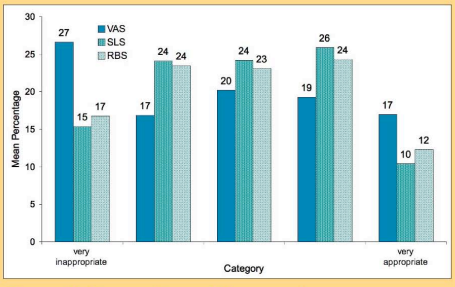


Figure 2. Comparison of frequencies

Conclusion

It makes a difference if VAS, SLS or RBS are used. Function is more important than appearance: even though SLS look more like VAS, they are used like RBS regarding data quality and distribution of values and we found no benefit in employing SLS. Confirming our previous finding on VAS' superior data level (Reips & Funke, in press), VAS again turn out to be the better scales. Present evidence shows them to be advantageous regarding dropout, missing data, and reliability of measurement. Drawing a line between discrete categories (as with SLS) is not enough. Superficial changes in appearance do not substitute for the power of a continuous measurement device like the VAS.

References

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review, 24*(2), 227-245.
Funke, F., & Reips, U.-D. (2007, June). *Improving data quality in Web surveys with visual analogue scales.* Paper presented at the second conference of the European Research Association, Prague (CZ).
Hartig, J., Jude, N., & Rauch, W. (2003). Entwicklung und Erprobung eines deutschen Big-Five-Fragebogens auf Basis des International Personality Item Pools (IPIP40) [Development and test of a German big five questionnaire based on the International Personality Item Pool (IPIP40)]]. *Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe Universität*, 2003/1. Frankfurt/Main.
Reips, U.-D., & Funke, F. (in press). Interval level measurement with visual analogue scales in Internet based research: VAS Generator. *Behavior Research Methods.*

39th annual meeting of the Society for Computers in Psychology SCiP 2009, November 19, Boston, MA

Making Small Effects Observable: Lowering Error with Visual Analogue Scales

Frederik Funke[1] & Ulf-Dietrich Reips[2]

Formatting Error

Error is an omnipresent opponent for those in charge of designing high-quality studies. Identifying and minimizing sources of error is a key interest of survey methodologists. The kind of error in focus here is *formatting error*. It is best described by combining two concepts: M*easurement error* – one part of the error of observation (e.g. Groves et al., 2004) – and the *question-answer process* (Schwarz & Oyserman, 2001). Formatting error happens when the exact extent of the true value to be measured cannot be communicated on the available rating scale.

The most frequent rating scales in self-administered studies are ordinal rating scales. If such a scale is used to measure on interval level, the expected formatting error for a single measurement decreases with the number of categories. It equals one fourth the distance between two categories. If seen as representing a latent continuum running from 0 to 100, for five-point scales it is as high as 6.2 percentage points (see Figure 1).

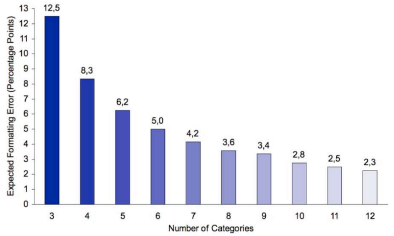


Figure 1. Cumulated formatting error with ordinal scales.

Visual Analogue Scales (VASs)

VASs are continuous graphical rating scales that produce data on the level of an interval scale (e.g. Reips & Funke, 2008). Respondents indicate the extent of the true value by placing a mark on a plain line. Every point on the line may serve as a possible answer. Thus, from a theoretical point of view, VASs have in contrast to all other ordinal rating scales an expected formatting error of zero (see Figure 2). VASs for local computerized assessment or for Web-based research can be created using the free Web service VAS Generator (http://vasgenerator.net).

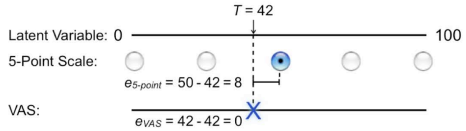


Figure 2. Formatting error *e* of true value *T* with 5-point scale and VAS.

Empirical Error

Study Design. In a between-subjects design 427 students taking part in a 40-item personality test were randomly assigned to five-point scales or to VASs of 250 pixels in length.

Results. We used type of rating scale as independent variable and the mean standard error as dependent variable. On a scale running from 0 to 100 the mean standard error with ordinal scales was M = 2.00 (SD = 0.40) and M = 1.76 (SD = 0.24) with VASs, $F(1,79)$ = 10.21, p = .002, η^2 = .12.

Conclusion

VASs produced significantly lower error in comparison to five-point ordinal scales. This is – according to our initial reasoning – because respondents can communicate their true values with greater precision when using VASs. Lower error leads to the confidence intervals being more narrow. Thus, VASs facilitate the detection of small effects that would otherwise not be observable. Due to higher precision of VASs any effects can be detected with smaller samples than needed with ordinal scales. On top of these advantages, data being on the level of an interval scale allow the greatest range of statistical procedures possible.

1: University of Tübingen, Germany, email@frederikfunke.de
2: Departamento de Psicología, Universidad de Deusto and IKERBASQUE (Basque Science Foundation), Spain, u.reips@ikerbasque.org

References

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology*. Hoboken, NJ: Wiley.

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet based research: VAS Generator. *Behavior Research Methods, 40,* 699–704.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22,* 127–160.

Funke, F., & Reips, U.-D. (2009, November). *Making small effects observable: Lowering error with visual analogue scales.* Poster session presented at the 39rd annual conference of the Society for Computers in Psychology (SCiP), Boston, MA (USA).

Funke, F., & Reips, U.-D. (2010a, April). *Better measurement with visual analogue scales: A Web experiment.* Poster session presented at the first joint meeting of the Experimental Psychology Society (EPS) and the Spanish Experimental Psychological Society (SEPEX), Granada (Spain).

1st joint meeting of the Experimental Psychology Society (EPS) and the Spanish Experimental Psychological Society (SEPEX)

Better Measurement with Visual Analogue Scales: A Web Experiment

Frederik Funke[1] & Ulf-Dietrich Reips[2,3]

[1]University of Tübingen, Germany; [2]University of Deusto, Spain; [3]IKERBASQUE (Basque Foundation for Science)

Problem: Formatting Error with Ordinal Scales

Whenever *ordinal rating scales* are used to measure *continuous latent variables* then the data are affected by *formatting error*. Formatting error leads to seriously problematic effects: Confidence intervals widen, statistical power is reduced, and small effects cannot be observed. *Figure 1* illustrates formatting error *e* occurring with a 5-point scale and a normally distributed data set for the middle category.

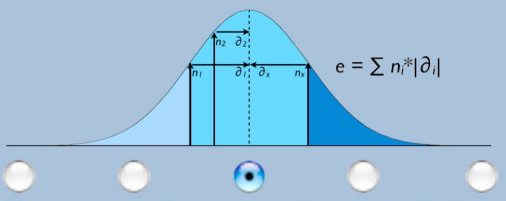
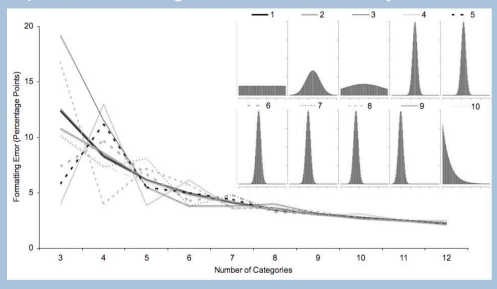


Figure 1. Formatting error *e* with normally distributed latent variable and 5-point scale.

Figure 2. Expected formatting error with 3 to 12 categories and differently distributed data sets; expected formatting error for VASs is always zero



The more options the better? We used simulated normally distributed data sets with different means and different dispersions to compute the expected formatting error. *Figure 2* shows the extent of formatting error not only to depend on the number of categories. It is determined by an interaction between the number of response options and the distribution of values. Thus, sometimes more categories result in more error. Particularly with a small number of response options the risk of running into substantial formatting error is evident, because the exact distribution of values can only be known after the data collection.

Solution: Visual Analogue Scales Tested in a Web Experiment

Visual analogue scales (VASs) are continuous rating scales (e.g. Hofmans & Theuns, 2008; Reips & Funke, 2008) that can easily be implemented in computerized data collection (*Figure 3*), e.g. with http://vasgenerator.net. From a theoretical point of view VASs have – regardless of the distribution of values – an expected formatting error of zero as there is a perfectly matching response option for every point on the continuous latent variable.

We conducted a *Web experiment* (e.g. Reips, 2007) with a within-subjects design where respondents (N = 265) were randomly assigned to VASs, 5-point, 7-point or to 9-point scales. We took the mean *standard error of the mean* (*SEM*) as a proxy for the formatting error for each of 58 items in a questionnaire.

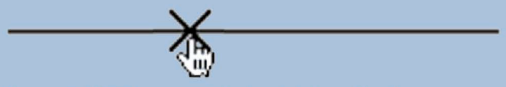


Figure 3. VAS for computerized data collection.

We found *SEM* with VASs (M = 3.47, SD = 0.40) to be lower in comparison to 5-point scales (M = 3.86, SD = 0.49) , $F(1, 115) = 21.55$, $p < .001$, $\eta^2 = .16$, to 7-point scales (M = 3.69, SD = 0.39) , $F(1, 115) = 8.95$, $p = .003$, $\eta^2 = .07$, but not statistically significantly lower in comparison to 9-point scales (M = 3.56, SD = 0.41), $F(1, 115) = 1.56$, $p = .215$.

Conclusion and Recommendation

VASs should be taken advantage of for measurement of continuous latent variables, resulting in *less error* and *more statistical power* in comparison to ordinal scales. Benefits: (1) small effects can be observed or – if larger effects are to be measured – sample sizes can be reduced; (2) data obtained with VASs rather than with ordinal scales can be recoded more flexibly, thus (3) allowing many more statistical analyses than with data from ordinal scales.

References

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring visual analogue scales. *British Journal of Mathematical and Statistical Psychology, 61*, 401–413.

Reips, U.-D. (2007). The methodology of Internet-based experiments. In A. N. Joinson, K. Y. A. McKenna, Postems, T., & Reips,, U.-D. (eds.), *The Oxford handbook of Internet psychology* (pp. 373–390). Oxford: Oxford University Press.

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet based research: VAS Generator. *Behavior Research Methods, 40*, 699–704.

Funke, F., & Reips, U.-D. (2010b, May). *Increase statistical power with visual analogue scales.* Poster session presented at the 12th annual General Online Research (G.O.R.) conference of the German Society for Online Research (D.G.O.F.), Pforzheim (Germany).

GOR 10

General Online Research 10, May 26–28, 2010, Pforzheim University, Germany

Increase Statistical Power With Visual Analogue Scales

Frederik Funke[1] & Ulf-Dietrich Reips[2,3]

[1]University of Tübingen, Germany, http://www.frederikfunke.de, email@frederikfunke.de
[2]University of Deusto, Spain, http://iscience.deusto.es; [3]IKERBASQUE (Basque Foundation for Science)

Formatting Error

… is the difference between true score and reported value; it occurs when ordinal rating scales are used to measure continuous latent variables (see Figure 1)

… lowers statistical power, raises the needed number of cases, and makes small effect sizes unobservable

– expected formatting error with visual analogue scales (VASs; e.g. Couper et al., 2006) is zero

Web Experiment

– questionnaire: 58 items on 29 Web pages

– between-subjects design: randomization to 5-point, 7-point, 9-point scales or to VASs (see Figure 2); *N* = 265

– computation of standard error of the mean (*SEM*) as indirect measure of formatting error

? Does number of response options empirically affect formatting error as theorized?

? Does response scale affect dropout and response time?

Figure 2. VAS in Web survey.

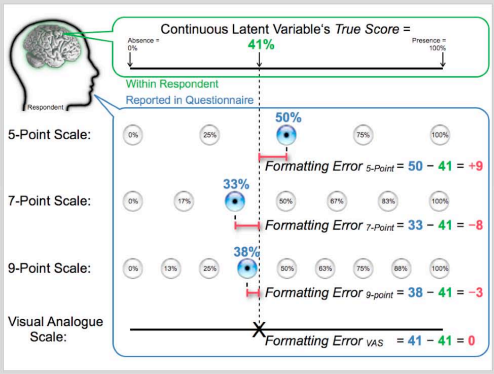


Figure 1. Formatting error occurring when measuring continuous latent variables with ordinal scales and with VASs.

Results

Scale	Dropout	Response time in min (SD)	Mean *SEM* (*SD*)
5-point	38%	9.0 (3.3)	3.9 (0.5)
7-point	20%	8.4 (3.8)	3.7 (0.4)
9-point	42%	8.6 (2.9)	3.6 (0.4)
VAS	24%	8.3 (2.8)	3.5 (0.4)

– difference in *SEM* between VAS and …
… 5-point: $F(1, 115) = 21.55$, $p < .001$, $\eta^2 = .16$
… 7-point: $F(1, 115) = 8.95$, $p = .003$, $\eta^2 = .07$
… 9-point: $F(1, 115) = 1.56$, $p = .215$, $\eta^2 = .01$

Conclusions

– the more response options the lower is formatting error; ordinal scales with *moderate* number of options perform better regarding dropout and response time

– only VASs combine low formatting error (*SEM*) and low dropout rate

– measurement with VASs leads to *less error*, thus *smaller confidence intervals* and *more statistical power:*
– *small effects* can be observed, the risk of making a *Type II error* is lowered
– *sample sizes* can be reduced if *large effects* are to be measured

– additionally, data from VASs allow more statistical analyses (Hofmans & Theuns, 2008; Myles & Urquhart, 2005; Reips & Funke, 2008) than data from ordinal scales and can be recoded and analyzed more flexibly

– consider using the free Web service http://vasgenerator.net to build VASs

References

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review, 24,* 227–245.

Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring visual analogue scales. *British Journal of Mathematical and Statistical Psychology, 61,* 401–413.

Myles P. S., & Urquhart N. (2005). The linearity of the visual analogue scale in patients with severe acute pain. *Anaesthesia and Intensive Care, 33,* 54–58.

Reips, U.-D., & Funke, F. (2008). Interval level measurement with visual analogue scales in Internet based research: VAS Generator. *Behavior Research Methods, 40,* 699–704.

Erklärung

Declaration

Erklärung

Hiermit bestätige ich, dass ich diese zur Promotion eingereichte Arbeit selbständig verfasst habe. Es wurden nur die angegebenen Quellen und Hilfsmittel benutzt und alle übernommenen Zitate wurden als solche gekennzeichnet. Die vorgelegte Dissertation ist bisher weder ganz noch teilweise als Dissertation oder sonstige Prüfungsarbeit eingereicht worden.

Declaration

I hereby declare that the work submitted in this dissertation is the result of my own investigation. The work is original except where indicated by citations and references. No part of this thesis has been submitted as dissertation or for any other degree.

Kassel, den 20. August 2010

_____
Frederik Funke